Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

Diplom-Physiker:	Sotirios Ziagos
born in:	Plettenberg, Germany

Oral examination:

A novel 4-valued discrete network for the analysis of signalling events integrating protein interaction and gene expression data applied on hematopoietic differentiation

Referees: Prof. Dr. Stefan Wölfl
PD Dr. Reinhard Guthke

English Summary

Signaltransduction is an important process, that is crucial for the interplay of cells in multicellular organisms and controls among other things cellular proliferation and cellular differentiation. The complexity of signalling networks and the absence of comprehensive and accurate quantitative data make a precise mathematical description of such networks difficult. This work deals with the questions, if and how predications about signalling processes for specific cellular models can be made with the aid of interaction databases and gene expression profiles.

Starting from molecular interaction-databases complex networks are generated comprising all known cellular signalling paths. But, not all of them play a role in a specific cellular model. The specific adaptation of the networks to a specific cellular system is achieved by additional integration of data obtained from gene expression profiles. Thereby, 1. the information about expression or non-expression of a gene and 2. the change of gene expression in time (increase, decrease or unchanged) is used.

1. Information about the expression is included in the calculations, by removing unexpressed elements from the network and thus changing its topology.

2. Chronological changes in expression are integrated in the estimations by a four-valued, discrete network containing statistical elements, that was developed in this work. This network is able to deal with contradicting data and pays attention to the importance of feedback loops for differentiation processes.

This enables an extraction of subnetworks having a dynamic that approximates the measured data best. Hematopoietic processes, playing among other things a role in development of leukaemias, were examined with a sensitivity analysis. The analysis revealed central molecules of EpoR induced hematopoesis, like AKT, PIP3, Pi3K, Lyn, ERK.

Deutsche Zusammenfassung

Signaltransduktion ist ein wichtiger Prozess, der für das Zusammenspiel der Zellen in mehrzelligen Organismen essentiell ist und u.a. Zellproliferation sowie Zelldifferenzierung kontrolliert. Die Komplexität von Signalnetzwerken und das Fehlen von umfassenden und genauen quantitativen Daten gestalten eine exakte mathematische Beschreibung solcher Netzwerke schwierig. Diese Arbeit befasst sich mit der Frage, ob und wie mit Hilfe von Interaktionsdatenbanken und Genexpressionsprofilen Aussagen über Signalprozesse für spezifische Zellmodelle getroffen werden können.

Ausgehend von Interaktionsdatenbanken werden komplexe Netzwerke generiert, die alle bekannten Signalwege in den Zellen zusammenfassen. Diese spielen jedoch nicht alle in einem spezifischen Zellmodell eine Rolle. Die spezifische Anpassung der Netzwerke an ein bestimmtes zelluläres System erfolgt durch zusätzliche Integration von Daten aus Genexpressionsprofilen. Dabei werden 1. die Information über Expression bzw. Nicht-expression eines Genes, sowie 2. die zeitliche Veränderung der Expression von Genen (Zunahme, Abnahme oder unveränderte Expression) verwendet.

1. Informationen zur Expression werden durch ein Entfernen der nichtexprimierten Elemente aus dem Netzwerk berücksichtigt und verändern somit die Topologie des Netzwerkes.

2. Zeitliche Veränderungen der Expression werden durch ein in dieser Arbeit entwickeltes vierwertiges, diskretes Berechnungsverfahren, welches auch statistische Elemente enthält, mit Widersprüchen zurechtkommt und die Wichtigkeit von Rückkoppelung bei Differenzierungsprozessen beachtet, berücksichtigt.

Dies ermöglicht eine Extraktion von Teilnetzwerken, deren Dynamik den gemessenen Daten am besten entspricht. Mit Hilfe einer Sensitivitätsanalyse wurden Prozesse der Hematopoese untersucht, die u.a. bei der Entstehung von Leukemien eine Rolle spielen. Dabei konnten zentrale Moleküle für die EpoR induzierte Hematopoese, wie beispielsweise AKT, PIP3, Pi3K, Lyn, ERK identifiziert werden.

Contents

Chapter 1

Introduction

Chronic myelogenous leukaemia (CML) is a myeloproliferative malignancy with clonal expansion of transformed primitive hematopoietic progenitor cells [1]. It represents 7–20% of all leukemia cases [2] and has an incidence of 1 to 2 cases per 100000 population [2, 3]. In almost all CML cases (90%) a chromosomal abnormality called Philadelphia chromosome (Ph) is found [3, 4]. The abnormality is a reciprocal translocation between the long arms of chromosomes 9 and 22, which leads to the generation of the BCR-ABL fusion oncogene and its protein product, the constitutively active BCR-ABL tyrosine kinase [5, 6]. The abnormal function of this kinase leads to dysfunctional regulation of cell growth and survival and therefore is sufficient to cause leukaemia [7, 8, 9].

CML can be treated in several ways [4, 3]:

Allogeneic Stem Cell Transplantation: This therapy is available to only 40% of the patients. It is the only available therapy with known curative potential. Stem Cell Transplantation has a treatment related mortality between 20 to 40% [10].

Chemotherapy: Busulfan is used to control haematologic variables and thus to reduce the symptoms of CML but it has serious side effects. Another chemical used in chemotherapy is hydroxyurea, a cell cycle specific inhibitor of DNA synthesis. Like Busulfan it allows haematologic control but has less side effects. Both chemicals are applied orally. They can neither produce cytogenic remission, nor delay death, their effects are mainly palliative.

Interferon-α: IFN can lead to haematologic and cytogenetic remission. Its effectiveness is increased when it is combined with AraC. The survival time of patients is increased by this therapy and in 5%–10% of cases the Philadelphia Chromosome disappears completely. The mechanism of IFN in CML is poorly understood. IFN therapy is in the long run not tolerated by 10%–25% of patients.

Imatinib: Imatinib is a novel agent that specifically inhibits the BCR-ABL tyrosine kinase. Long-term survival data are not available for imatinib-treated patients, but data obtained till now suggest that imatinib therapy is more effective and better tolerated than IFN/Ara-C therapy.

Except Stem Cell Transplantation all therapies work by changing cellular signalling. This is especially the case for IFN which is targeting the IFN-Receptor and Imatinib which inhibits specifically the BCR-ABL tyrosine kinase. Thus a deeper understanding of function of current agents and discovery of new targets require methods to understand the processes on the signalling level.

1.1 Hematopoesis

Hematopoiesis (from Greek haima for blood and poiein to make) is the development of blood cells. Prenatally hematopoiesis occurs in the yolk sack, later in liver and spleen. Next, the bone marrow becomes responsible for hematopoiesis and hematopoiesis in liver and spleen ceases. In normal adults hematopoiesis occurs in marrow and lymphatic tissues. Although hematopoietic stem cells are the best studied stem cells in mammals up to date, [11, 12] some questions still are unanswered or are discussed controversially [11, 13]. Most studies were performed on mice but there are big parallels between murine and human hematopoietic systems [13].

1.1.1 Differentiation processes

All mature blood cells from *T Cells* to *Neutrophils* and *platelets* are generated from a small number of hematopoietic stem cells (HSC) and progenitors [11, 14]. In mice a single HSC is sufficient to rebuild the entire hematopoietic system of an animal [15]. This shows that HSCs can generate a large number of cells and that they are able to create any cell-type belonging to the hematopoietic system.

HSCs produce Common Lymphoid Progenitors (CLP) and Common Multipotent Progenitors (CMP) [16, 17]. This is an irreversible multistep process [14]. CLPs and CMPs cannot produce all cell types of the hematopoietic system any more. CLPs can generate only B, T and NK cells whereas CMPs have the ability to produce just red cells, platelets, granulocytes and monocytes [16, 17]. Figure 1.1 shows natural chemokines that induce the development of certain progenitors.

For this work it is important to look closer at the differentiation stages leading to the development of erythrocytes (see also figure 1.1): Hematopoietic stem cells (HSC) are stimulated with TPO and thus differentiate into common multipotent progenitors (CMP). Stimulation of CMP with EPO produces megakaryocyte erythrocyte progenitors (MEP) which can further differentiate to erythroblasts and erythrocytes if they are further stimulated with EPO. I would like to note that the interplay of cytokines leading to erythrocytes in reality is much more complex [18].

In addition to their ability to differentiate into mature blood cells, HSC and other hematopoietic cells are able to remain quiescent, proliferate, differentiate, self-renew or undergo apoptosis with these cellular processes depending as well on intrinsic genetic processes as on external factors, like cytokines, chemokines and others [20, 21, 22].

Hematopoietic cell-fate-decision is explained by the instructional model:

instructional model: The instructional model assumes that the environment plays the primary role in determining the fate of HSCs and pro-

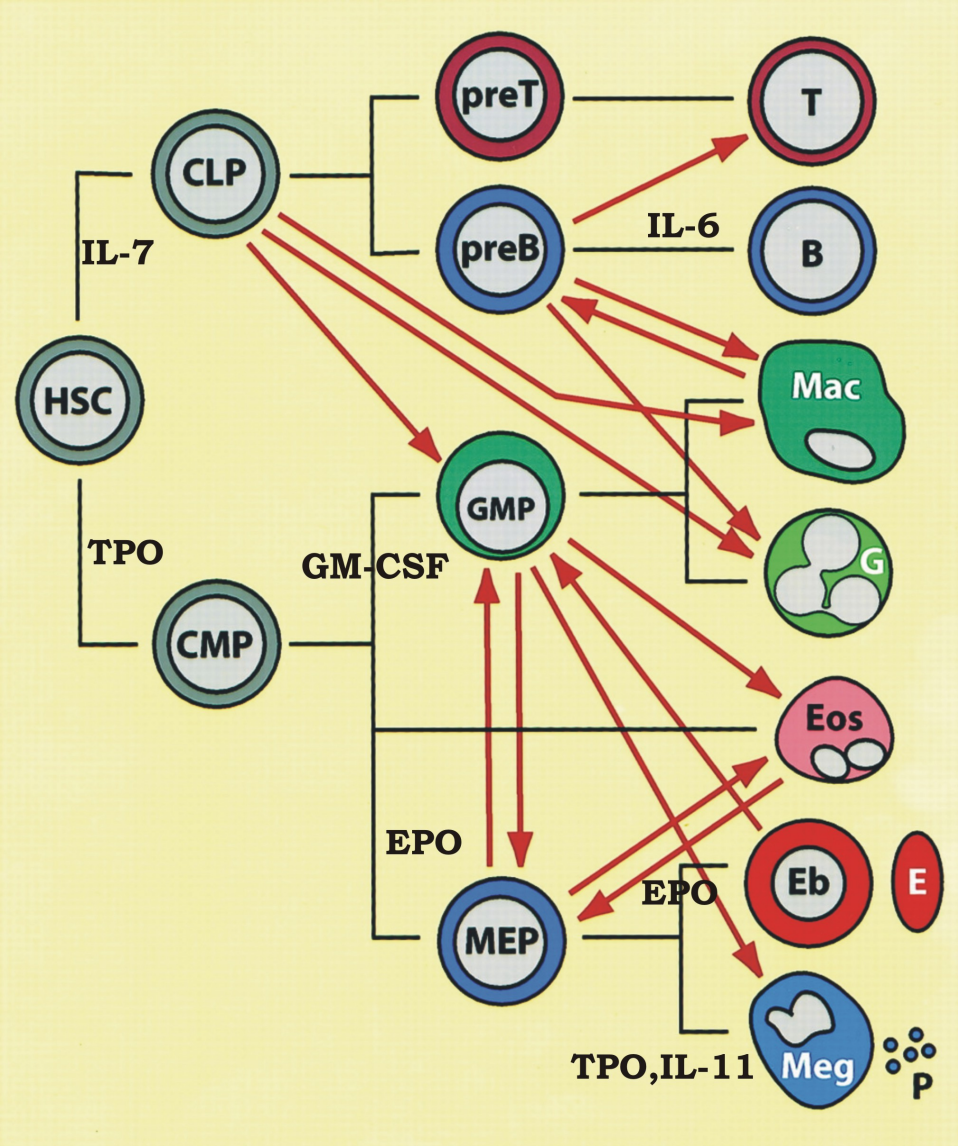


Figure 1.1: Normal hematopoietic differentiation and lineage switches
The black lines indicate normal lineage relationships and the responsible ligands , the thick grey arrows represent induced switches. (These do not necessarily imply direct transitions.) HSC indicates hematopoietic stem cell; CLP, common lymphoid progenitor; CMP, common multipotent progenitor; GMP, granulocyte/macrophage progenitor; MEP, megakaryocyte erythrocyte progenitor; T, T lymphocyte; B, B lymphocyte; Mac, macrophage; G, neutrophil granulocyte; Eos, eosinophil, Eb, erythroblast; E, erythrocyte; Meg, megakaryocyte; P, platelet. Mast cells, NK cells, and dendritic cells have been omitted from the scheme and placement of eosinophils is speculative. Note that most of the switches were observed with transformed cell lines in culture and that the cell type designations (both before and after the switch) may not accurately reflect the phenotype of the normal counterparts. Modified from [19]

genitors, thus environmental factors determine lineage choice and cellular development [23].

1.1.2 Lineage switching

In the last few years a series of studies has indicated that hematopoiesis is not an irreversible hierarchical process and that cells that were believed to be developmentally restricted can recover their ability to commit to other lineages [27, 19, 28]. Figure 1.1 shows a summary of observed changes. The cells showing this behaviour are mainly transformed and altered cells but also in normal cells, at least in culture, transdifferentiation has been observed [29]. It is unknown whether transdifferentiation involves retrodifferentiation to an earlier progenitor or if transdifferentiation occurs directly between different lineages. Despite the fact that some cells have the ability to retrodifferentiate, most lineage switches most likely do not involve retrodifferentiation [30, 19]. Lineage switches do not occur *in vivo* or occur in very low frequencies under normal conditions [19].

Aforementioned lineage switching has been studied intensively in the past years. Many studies which were carried out that show, how enforced transcription factor expression results in lineage plasticity [31, 32, 33, 34, 35, 36, 37, 38].

1.1.3 CML

As already mentioned in the introduction CML is caused by the BCR-Abl fusion oncogene. Three main BCR-Abl isoforms are known $p190^{Bcr-Abl}$, $p210^{Bcr-Abl}$ and $p230^{Bcr-Abl}$, each leading to a different disease phenotype [39]. Classical CML is associated with $p210^{Bcr-Abl}$ [40].

The $p210^{Bcr-Abl}$ isoform leads to an increase in transformed granulocytic and megakaryocytic lineages. The majority of cells is arrested at intermediate stages of differentiation, but a small proportion of cells become mature neutrophils, eosinophils and basophils [41].

In addition to this rather macroscopic observation transformed cells also have some altered characteristics on the cellular level [42], which are known.

Activation of Mitogenic Signalling Pathways: In Bcr-Abl transfected cell lines increased activity of pathways having mitogenic potential is observed. For example Ras, MAP Kinase, PI-3/Akt Kinase, Myc and STAT5 pathways get influenced by Bcr-Abl [43, 42].

Inhibition of Apoptosis: It is discussed controversially whether Bcr-Abl inhibits apoptosis or not *in vivo*. However, transfection with exogenous BCR-ABL leads to immortality against growth factor withdrawal and increased resistance to apoptosis, induced by DNA-damaging agents.

Proteasomal Degradation: Bcr-Abl can promote the degradation of proteins via the ubiquitin-proteasomal pathway.

Genomic instability: CML evolves in 3 distinct clinical stages: chronic, and accelerated phases and blast crisis. This is possible because Bcr-Abl promotes secondary molecular and chromosomal changes, which with time increase malignancy of the illness. The reason for this effect might be the inhibition of apoptosis or the dysregulation of DNA repair that are induced by Bcr-Abl.

Altered Cellular Adhesion: Healthy Hematopoietic progenitors are held in position by adhesion within the vicinity of cytokine-secreting cells, determining their fate (see previous sections). CML cells exhibit reduced adhesion and thus receive weaker regulatory signals.

1.2 Signal transduction – dynamic and general rules

> 'In biology, signal transduction is any process by which a cell converts one kind of signal or stimulus into an other. Processes referred to as signal transduction often involve a sequence of biochemical reactions inside the cell, which are carried out by enzymes and linked through second messengers. Such processes take place in as little time as a millisecond or as long as a few seconds. Slower processes are rarely referred to as signal transduction [44].'

Signal transduction occurs within a network, leading sometimes to unpredictable effects, here are some general principles and effects found in signalling networks.

Phosphorylation and dephosphorylation: It is possible to control enzymatic activity reversibly by phosphorylation and dephosphorylation [45]. Kinases phosphorylate their downstream molecules whereas phosphatases dephosphorylate them [45]. Kinases control primarily the amplitude of a signal and phosphatases its rate and duration [46].

Second messengers: Second messengers are non protein molecules used to relay a signal within a cell [47]. They can be synthesized/released and broken down again in specific reactions by enzymes [47]. Second messengers are used mainly in case a signal is communicated through extracellular membrane receptors [48]. Examples for secondary messengers are: cAMP, cGMP, IP_3

Ultrasensitivity and thresholds: Cascades with ultrasensitive stimulus response amplify signals above a threshold and reduce signals below in an all-or-none manner [49]. These could be very important for events being binary in nature [49]. This leads naturally to increased robustness against fluctuations [50]. Ultrasensitivity obtained by multisite-phosphorylation is very robust in contrast to zero-order-ultrasensitivity [49, 51]. Zero-order-ultrasensitivity is unlikely in most biological systems [51]. Threshold regulation is used in *Drosophila* to prevent inappropriate receptor tyrosine kinase responses to moderate signalling levels [53].

Random effects Two types of noise are distinguished: Intrinsic and extrinsic noise [54]. Intrinsic noise is based on the fluctuations in molecule number of a single component due to random birth and decay processes [55, 54, 56]. Extrinsic noise is the fluctuation imposed on a component by other network members by changes in reaction rates [54, 56, 57]. Noise in cells is generated primarily by low numbers of molecules. Regarding protein generation following rules apply [58, 59]:

1. Large fluctuations are observed at low transcription and high translation rates
2. Small fluctuations are observed at high transcription and low translation rates
3. Maximum external noise is observed at intermediate levels of expression

Noise is required for cell fate decisions and population heterogeneity in hematopoietic cells, too [57, 60]. Positive feedback can amplify the effects of noise and thus create a switch-like response [57, 61].

Noise can be decreased by negative feedback loops, redundancy, regulatory checkpoints, or stochastic focusing [62, 63, 58, 64, 65, 66, 67].

Feedback : A feedback loop is a circular chain of interactions. For any circuit each element exerts an indirect effect to itself, this effect can be either positive or negative. The indirect self regulation that each element exerts on itself, is the same for all elements of a feedback loop. This leads to the definition of the 'circuit sign'. The circuit sign depends on the number of negative interactions, if their number is odd, the circuit is negative, otherwise it is positive [68]. At least one positive feedback loop is required to generate multistationarity [68].

1.3 Computational approaches

In the past years many methods were used to understand and model cellular networks. Applied methods range from directed and undirected graphs [69] over Bayesian, Boolean [70, 71], rule-based [72] and generalized logical networks [73, 74] to stochastic networks [75] and networks using differential equations [76, 77, 78, 79].

Stochastic networks and networks based on differential equations require usually a vast amount of knowledge, therefore these networks are best applied on well studied small systems, like phages or on separable substructures of big systems.

Most models using discrete time can be treated as special cases of Dynamic Bayesian Networks (DBNs) [80]. Additionally, DBNs have some other beneficial features:

- They are well suited to deal with stochastic phenomena. This is important, because stochastic events seem to play a role in differentiation.
- They are able to deal with noisy data. This is relevant, because gene expression profiles are generally noisy.

- They are opening the way for likelihood maximization algorithms and thus can infer hidden parameters and handle missing data. In the human system it is very likely that data is not complete at present.

Because of all these reasons I believe that Bayesian networks are suited well to model differentiation of hematopoietic cells.

1.3.1 Bayesian networks

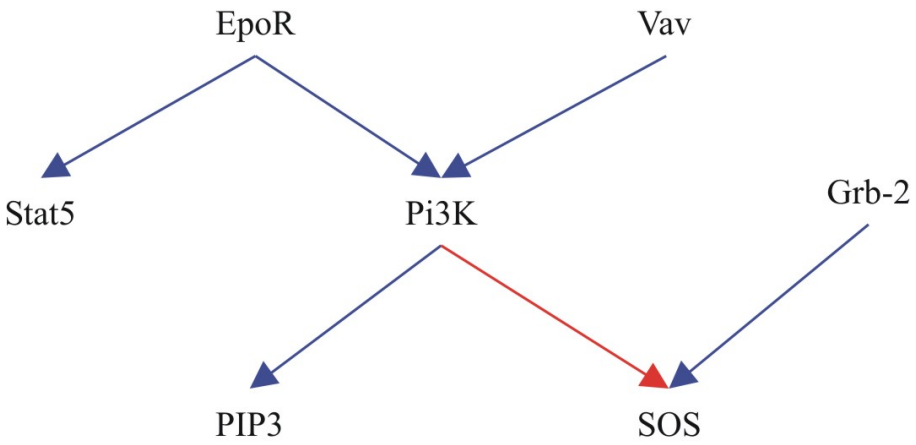


Figure 1.2: A simple network
This is an acyclic directed graph. The variables (nodes) represent states of molecules, the directed links represent interactions. Blue links represent activations and red ones inhibitions. The network is constructed from data found in the Transpath database (see section: 2.2.1)

Figure 1.2 shows some interactions that are found around the kinase Pi3K, in a biological database, building a causal network. The network consists of *variables* representing the status of molecule species and *directed links* between them symbolising interactions. A blue arrow indicates activation, a red arrow inhibition. The whole structure is called a *directed graph*. If there is a link from A to B, then B is called *child* of A and A is a *parent* of B. In Figure 1.2 EpoR is *parent* of Stat5 and Pi3K. Pi3K is *child* of EpoR and Vav.

Bayesian networks are used to calculate *conditional probabilities* of events. In the above example a conditional probability statement could be as follows: 'If EpoR (A) has an increased activity, then how big is the probability that Pi3Ks (B) activity is increased or decreased' . If it is assumed, that both molecules are expressed and Stat5 is not saturated etc., then the answer would be: With a probability of 1 Stat5 has also increased activity. The above statement can be expressed mathematically as: $P(B \mid A) = (1, 0)$,

$A = (1, 0)$, where (1,0) means that the probability for increased activity is 1 and for decreased 0. $P(B \mid A)$ can be translated into 'the Probability distribution over the possible states of B if A is known'. The *Fundamental Rule* for probability calculus is:

$$P(B \mid A)P(A) = P(A, B)$$

with $P(A, B)$ being the probability $A \wedge B$ that A and B happen together. This leads to the *Bayes Rule*[81] :

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

In the above example it was easy to calculate the probability distribution of Stat5, if EpoR is known, because in this simplified graph Stat5 has only EpoR as parent. Things become more complicated for SOS, because it has two contradicting parents, Pi3K and Grb-2, the conditional probability of SOS depends on the status of Pi3K as well as on the status of Grb-2 $P(SOS \mid Grb-2, Pi3K)$. The network in Figure 1.2 would require specification of the following probabilities to be complete: $P(EpoR)$, $P(Vav)$, $P(Stat5 \mid EpoR)$, $P(Pi3K \mid EpoR, Vav)$, $P(PIP3 \mid Pi3K)$, $P(SOS, Grb-2 \mid Pi3K)$ [81].

In principle, in Bayesian networks the correlations between variables in the data are used to calculate required probabilities (and thus the topology) by calculating the most probable model given the data using the Bayes Rule [82, 83]:

$$P(Model \mid Data) = \frac{P(Data \mid Model)P(Model)}{P(Data)}$$

with:

$$P(Data \mid Model) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!$$

where $P(Model \mid Data)$ is the probability, that the model is right, if the data is given. This is the value that has to be maximised.

$P(Data \mid Model)$ is basically the product of the probability of observing child nodes, i in a particular state, k if parents are in some state j [82].
n is the number of measured variables.
$q - i$ is the number of possible combinations of values for the parents of variable i.
r_i is equal to the number of states that are possible for variable i.
N_{ij} is the number of cases in the dataset where parents of the variable i are in state j.
N_{ijk} is the number of cases in the dataset, where variable i is in state k and its parents are in state j

1.3.2 Bayesian networks applied to signalling

Woolf *et al.* [82] describe how they used Bayesian networks to analyse cell fate decision in embryonic stem cells, a problem that is similar to the analysis of differentiation processes in hematopoietic stem cells.

Woolf *et al.* used

> '49 measurement of phosphorylation states of 28 cellular signalling proteins along with cell proliferation and differentiation responses, across a landscape of stimulatory extracellular cytokine cues [82]'

to learn their Bayesian network. In order to get enough data to learn their network Woolf *et al.* used a resampling and discretisation approach to convert a single observation into 1000.

They applied the following simplifications on their model:

1. Data taken at different time points are independent of each other
2. The underlying network can be represented as a directed acyclic graph
3. All nodes have less than four parents and/or children
4. Special nodes, called cue, (biologically these correspond to receptors) have no parents
5. Special nodes ('high-level, phenomenological outputs of a system'), called response, have no children

The second simplification is necessary to maintain logical consistency because till today no calculus has been developed that can cope with feedback cycles [82, 81, 84]. The other simplifications are required to limit the search space and deduce directions.

In their work Woolf *et al.* use the generated data to explore the search space of possible element-element connections in order to calculate the most probable network. The resulting network is represented as a directed graph that connects the 28 measured proteins (see figure 1.3), the used cues and the measured phenomenological outputs with each other. Therefore the network has 38 nodes. Woolf *et al.* needed approximately two weeks of computing time on a 64-processor IBM pSeries 655 server for network calculation.

The main biological benefits are that, given the expression data together with errors, a network explaining the data is derived automatically. This network could then be used to generate hypotheses helping to select drug targets [82].

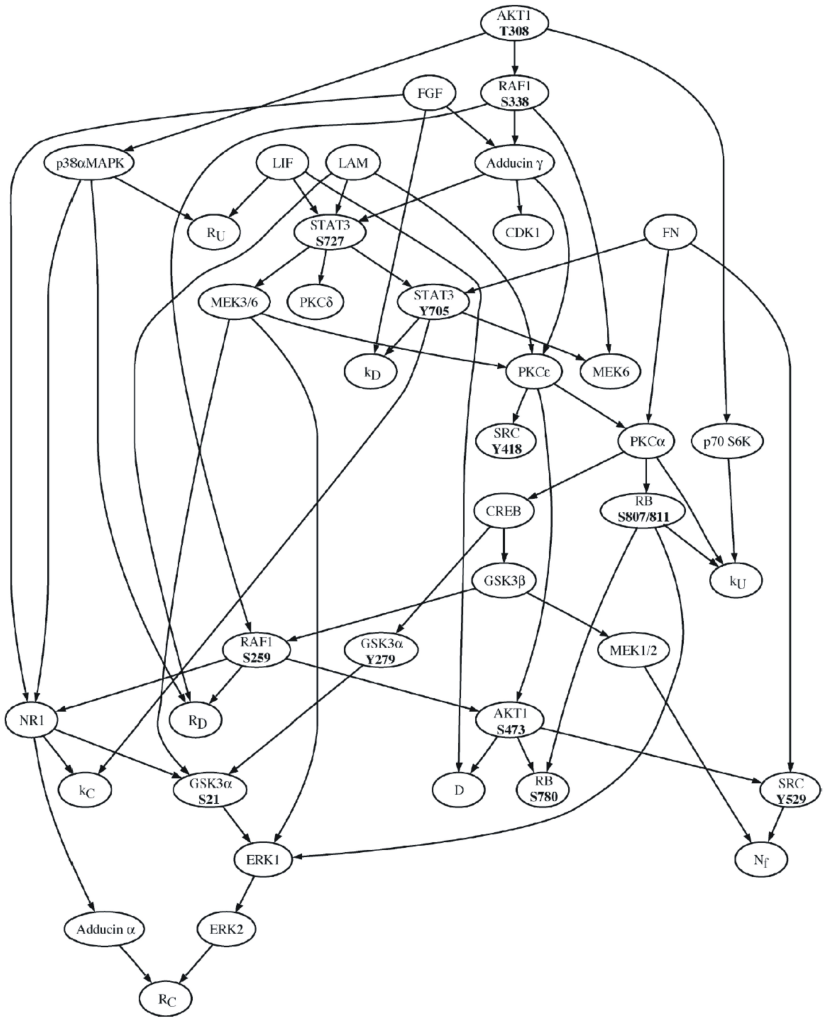


Figure 1.3: The network used by Woolf et al.
This is the network resulting from the analysis in [82]. The nodes are the 28 + 12 measured proteins, cues and 'high level, phenomenological outputs'. The arcs show the calculated interactions between them deduced from 49 measurement of phosphorylation states of the 28 proteins.

1.3.3 Bayesian networks and data connection

Bayesian approaches have also been used to infer gene networks [85, 86, 87] from gene expression data. However in this context it seems to be necessary to include additional biological information to obtain correct results [88, 89, 90]. Imoto *et al.* present in [90] a method that allows to integrate virtually any additional data to Bayesian networks.

In their model Imoto *et al.* use two parameters to find the optimal solution: The prior probability and the marginal likelihood of the data. They add the biological data into the prior probability of their model and use the marginal likelihood to measure the fitness of the model to the microarray data. The prior probability is calculated by assigning an 'interaction energy' to each known biological fact. Using these set values, calculations are performed recursively together with the expression data to calculate new *interaction energies*. Finally, this procedure leads to an optimal solution.

Imoto *et al.* have to constrain their search space, otherwise their calculations are intractable even for moderately sized networks [90]. Therefore they limit their search space by two extreme scenarios and assume that the real network is somewhere in between.

- The 1st extreme case is a cyclic network.
- The 2nd extreme case is a strictly hierarchical network.

Imoto *et al.* apply their method on *Saccharomyces cerevisiae* gene expression data. They restrict their analysis to downstream genes of 5 transcription factors. For the analysis they use gene expression data obtained by disrupting 100 genes.

The biological benefit is, that by adding biological knowledge into the network, more information from microarray data gets extracted leading to a more accurately estimated gene network [90].

1.3.4 Dynamic Bayesian Networks

The two mentioned methods avoid cycles and disregard temporal information of networks. Dynamic Bayesian networks are believed to be able to deal with temporal information and feedback loops [80, 91, 92].

In contrast to non-dynamic Bayesian networks, where $P(Data \mid Model)$ is calculated neglecting the time-order of the measurements, in Dynamic Bayesian Networks the time-order is not ignored, this allows to some extent to use time as an additional dimension to transform cyclic networks into acyclic ones [80]. The incorporation of time is obtained by a statistical analysis of data across different time slices. This means for example that in a Dynamic Bayesian Network a check is done, whether a relation between A and B exists for A(t=1)and B(t=1), A(t=1) and B(t=2),A(t=2)

and B(t=3),A(t=1) and B(t=3) etc.. In contrast to this in non-dynamic Bayesian networks only relations in data of the kind A(t=1)and B(t=1), A(t=2) and B(t=2) etc. are considered.

Perrin *at al.* and Zou *et al.* analysed gene networks with dynamic Bayesian networks [91, 92]. Perrin *et al.* examine *S.O.S DNA Repair* in *E.coli*, they use a time series of 8 genes to learn their model. The time series consists of 50 measurement points performed in 6 minutes intervals. Zou *et al.* model yeast cell cycle with 16 time points done in time intervals of 10 minutes. The network is restricted to 9 transcription factors and 116 genes, that are regulated by these transcription factors.

Using prior information the model of Zou *et al.* is able to identify 46 correct relationships and 1 misdirected. Without prior knowledge their model identifies 17 relationships and 3 misdirected ones. Zou *et al.* do not mention in their publication how many relationships are missed but if every gene is regulated by a single transcription factor then there should be 116 interactions in the network. Thus most likely even with prior knowledge more than half of the interactions are missed, although the examined network is comparatively small.

Zou *et al.* claim ,that the methods predicts gene-gene relationships more accurate than traditional Bayesian Network methods [92].

1.4 Aim of the work

The described works of current approaches have some common properties.

- The data and Bayes rule is mainly used to deduce the complete topology of the network. This is done even in the work of Imoto *et al.*, where prior biological knowledge is used to calculate prior probabilities of the models.
- The treated systems are rather small. Either they deal only with a fixed number of pre-selected molecules or with a handful transcription factors and their associated genes and still require a lot of data to calculate results.
- The models avoid cyclic structures either by not allowing them or by making them acyclic, using time, despite the fact that feedbacks are considered to be very important for differentiation processes [93, 94].
- Most models deal with gene networks and disregard interactions on protein-protein level. The only exception is the work by Woolf *et al.*, where phosphorylation events are measured to build a signalling network [82]

For analysis of Epo induced differentiation in stemcells affymetrix data for four different time points are available. The chip is able to measure approximately 33000 genes. The number of genes and thus the number of possible networks is too big to be feasible with above mentioned methods. Using the transpath database and restricting the network solely to elements being downstream of the erythropoietin receptor reduces the network to a network of approx. 1000 elements. This network is still more than eight times larger than the biggest networks described above and additionally has a much more complex structure (see fig. 1.4). In mentioned works the number of different measured points varied from 16 [92] to 100 [90], which is 4 to 25 times more data points than those that are available for the task at hand. All in all, the methods above require way too many data and therefore cannot be applied to the task at hand without modifications.

Altogether the main problem is the enormous search space that would be required if the structure would be deduced from the data. In this work I will present a method that can be used to describe the regulatory network underlying erythroid differentiation despite a low number of measurement points. A model for this requires the following characteristics:

- It does not demand many data points even for big networks to produce results.
- Because positive feedback cycles seem to be important [73] it must be able to deal with positive feedback cycles.
- It should not assume in contrast to previous methods, that the relation between elements remains the same all the time.

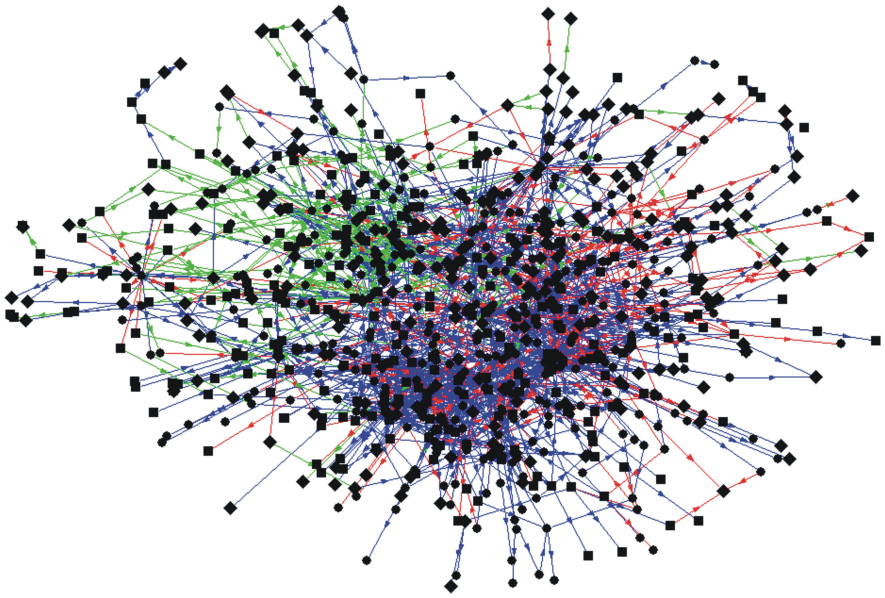


Figure 1.4: EpoR-Network:

This is the network downstream of the EpoR as found in Transpath. Red arrows represent inhibiting interactions, blue arrows activating reactions and green arrows interactions between genes and proteins and vice versa. The network consists out of 983 nodes.

Chapter 2

Formulation of a discrete model that describes aspects of signalling dynamic

The main problem with large networks arises from the fact, that the learning process from data is a NP-complete problem, this means that computation time increases exponentially with problem size [95]. Therefore the only way to deal with large networks is to restrict the search space from the beginning. As prior biological knowledge is available it is logical to use this knowledge to restrict the system. For this purpose the Transpath database was used, a database containing many protein-protein interactions on the signalling level. The knowledge of the database can be represented as a directed graph (see figure 1.4) having proteins, secondary messengers and other elements as nodes and their interactions as edges. The edges can be of different types, e.g. 'activating', 'inhibiting'.

Bayesian networks and biochemical directed graphs represent different concepts [82], therefore it is necessary to develop a method that allows to use biological networks together with Bayesian networks. A way how this can be done is shown in the following sections.

2.1 Dynamic conditions during *hematopoietic* differentiation

Hematopoesis appears to be, at least partly, a probabilistic process: At low levels of multiple cytokines basal hematopoiesis occurs randomly. Higher doses of cytokines force the cells to differentiate into a specific way [96]. It is likely that basal hematopoiesis is based on low transcription and high translation rates of proteins, giving rise to stochastic processes. It is likely that the m-RNAs of these proteins cannot be detected by Affymetrix chips

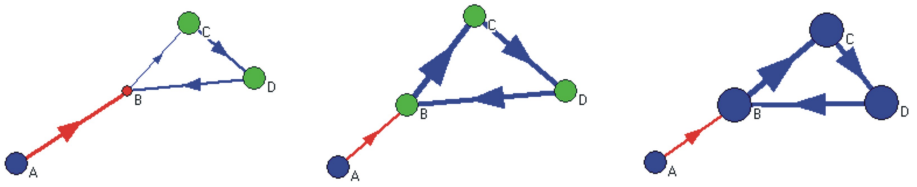
and therefore might be marked as absent. By chance these proteins exceed a threshold, this leads to a self-sustained increase or stabilisation at higher level of transcription, and thus to differentiation in a certain direction.

It is known that protein concentration is sensitive to external signals at intermediate expression levels. This leads to the conclusion that proteins being responsible for above mentioned bistability could also be controlled well by cytokines at intermediate expression levels, thus levels below the threshold. If a protein concentration reaches an intermediate level but is not favoured by the cytokine, then it is unlikely that it will exceed the threshold, on the other hand a favoured protein will be supported to exceed the threshold.

This model also explains why a cell if it reaches a certain state of maturation is developmentally inhibited, despite the presence of cytokines. If the intermediate expression level cannot be obtained by fluctuation, then cytokine control is less effective.

To overcome probabilistic behaviour and thus establish a sustained value above/below a threshold a higher/lower production of the proteins must be somehow established in a stable way. There are multiple ways to achieve this:

1. positive feedback loops self-activation, also multistep self-activation get active if the threshold is exceeded.



(a) A is inhibiting B therefore the positive cycle cannot be established

(b) By chance B exceeds a threshold value, strengthening the feedback cycle

(c) The feedback cycle is so strong that it prevails despite inhibition

Figure 2.1: Multistep self-activation

2. inhibition of an inhibiting process gets active, if the threshold is exceeded

3. if the threshold is not reached, then a negative feedback loop gets active (for a protein that is present at high concentrations but has to be removed)

4. if the threshold is not reached, then an antagonist protein gets activated (for a protein that is present at high concentrations but has to be removed)

5. sustained positive input of a cytokine keeping the value above a threshold

6. sustained negative input of a cytokine keeping the value below a threshold

All scenarios have one feature in common: they contain linear and/or cyclic self perpetuating structures. Assuming that hematopoiesis is based on positive feedback loops, then it is sufficient to examine these self sustaining structures.

2.2 Formal treatment in a discrete network

Assuming that the Transpath database is complete, it is necessary to identify all possible self sustaining structure combinations being allowed in the database and afterwards find those that match the measured data best. In the subsections below it is described how to solve these problems.

2.2.1 Introduction into the Transpath database

The Transpath database contains information about molecules which play a role in signal transduction and interactions among them. It focuses on signalling cascades that aim at transcription factors and hence change expression. The database Transfac contains information about interactions between Transcription factors and genes. Both databases used together make it possible to investigate the interplay between events on the protein level and events on the genetic level. Transpath version 5.3 has 19023 entries about molecules, 4989 entries about genes and contains 23062 reactions [97]. The entries are manually curated by experts and hence are reliable.

In the database molecules are connected by reactions with each other. The reactions connecting them have properties such as 'activation', 'inhibition', 'unknown' and others. In this work the required information was extracted from the database and was modified to satisfy the needs for modelling. All indirect reactions as well as all reactions that were neither 'activations' nor 'inhibitions' were disregarded, this lead to a directed graph. As the database distinguishes between orthologues and organism specific molecules and organisms and the organism specific reactions are not complete, all molecules were substituted by their orthologue to get a network as complete as possible.

2.2.2 Information content of Affymetrix data

Gene expression results are generally believed not to be very precise, therefore the information gained by this data is best treated as discrete entity. Standard analysis with Microarray Suite from Affymetrix yields easily the

information whether gene expression of a gene 'increases' or 'decreases' with time and whether a certain gene is 'absent' or 'present' at a certain time point. As Transpath combined with Transfac contains also information on how signalling networks influence gene expression, it is obvious to use this data to fit calculations. But the results obtained by Microarray Suite analysis can be used in an additional way.

As highly regulated proteins, thus proteins that show high difference in expression in affymetrix chips, have a high correlation between m-RNA concentration and concentration of synthesised proteins (r=0.89) [98], changes in expression profile can be used to estimate changes in protein concentrations. This is only possible for genes that are controlled at the transcriptional level. So one might conclude that 'Protein a has a higher concentration, because its gene A is higher expressed', but the conclusion 'Concentration of Protein b remains constant, because expression of Gene B is unchanged' is not allowed, because concentration of b might change, if b is controlled primarily on the translational and/or post-translational level. Assuming that higher protein concentration correlates with increased activity of a protein, with some caution gene expression results can also be used to estimate events, which occur on the signalling level.

These discrete properties can be used to infer probabilities, if the directed graph from Transpath is known.

2.2.3 Probability inference on directed graphs

Acyclic graphs

The following abbreviations and symbols will be further supplemented and used throughout this work:

- For the data obtained by Affymetrix measurements abbreviations are defined as follows:

  IncrExpr : Affymetrix Suite 5.0 indicates, that expression of the gene is increased.

  DecrExpr : Affymetrix Suite 5.0 indicates, that expression of the gene is decreased.

  ABSENT : Affymetrix Suite 5.0 indicates, that gene is not expressed. If a protein is not expressed, then it is not able to interact, therefore molecules produced by these genes were removed from the network, if not at least one homologue gene was expressed.

  PRESENT : Affymetrix Suite 5.0 indicates, that gene is expressed. Proteins whose genes are expressed remain in the network and are able to perform all their downstream interactions.

- For the state of the molecules and genes calculated using Transpath, the model and Affymetrix data, the used symbolism is given as follows. Only genes and molecules that were not removed from the network can be in one of the described states:

  0 : The activity of the molecule is not changed.

  Rule 1 (Calculating '0') *If all parents of the molecule are 'ABSENT' or have the value '0' themselves and all genes producing the molecule are neither 'IncrExpr' nor 'DecrExpr', then the molecule receives the value '0'.*

  + : The activity of the molecule in the network gets increased or is already maximal (e.g. saturation).

  Rule 2 (Initialise '+') *Proteins encoded by genes with 'IncrExpr' and molecules which are known to have increased activity (e.g. by measurements) receive '+' as initial value.*

  Rule 3 (Calculate '+') *'+' is calculated as value for a molecule using the topology of Transpath, if all activating parents have the value '+' or '0' and all inhibiting parents the value '-' or '0' and at least one parent is either '+' or '-'.*

  - : The activity of the molecule in the network gets decreased or is already zero (e.g. complete inhibition).

  Rule 4 (Initialise '-') *Proteins encoded by genes with 'DecrExpr' and molecules which are known to have reduced activity receive '-' as initial value.*

  Rule 5 (Calculate '-') *'-' is calculated as value for a molecule using the topology of Transpath, if all activating parents have the value '-' or '0' and all inhibiting parents the value '+' or '0' and at least one parent is either '+' or '-'.*

Applying these rules to the network shown in figure 2.2 the following conclusions can be drawn. At the beginning all molecules in the network are in the state '0'. The network is consistent with all rules, but does not contain much information.

The experiments are performed with a medium, that contains Epo, which is activating EpoR, thus EpoR is initialised with '+' (rule 2). By applying rules 1, 3, 5 to the network one obtains EpoR= '+', Vav= '0', Stat5= '+', Pi3K= '+', PIP3= '+', SOS= '-', Grb-2= '0'. It is possible to apply all the rules easily and straightforward.

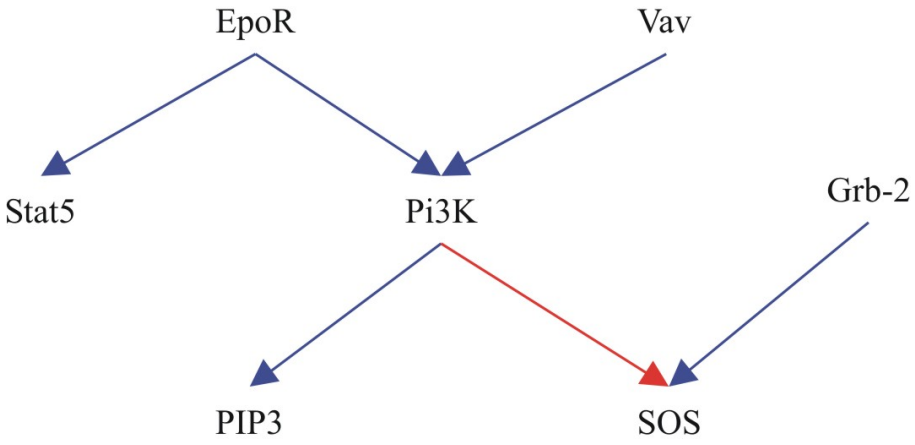


Figure 2.2: Previous network example
This is an acyclic directed graph. The variables (nodes) represent states of molecules, the directed links represent interactions. Blue links represent activations and red ones inhibitions. The network is constructed from data found in the Transpath database (see section: 2.2.1)

This changes, if contradictions are introduced into the network. As an example let the states of all molecules of the network again be '0' and assume that measurements show 'IncrExpr' for EpoR and 'DecrExpr' for Vav, applying rules 2, 4 on the network one gets EpoR= '+' and Vav= '-'. Stat5= '+' (rule 3), but the value of Pi3K cannot be calculated with rules 1–5. Pi3K has in this example a '+' activating parent and a '-' inhibiting parent the available knowledge is not sufficient to decide which effect is stronger and thus it is not possible to infer what happens to Pi3K. Activity of Pi3K could increase, decrease or even not change at all. Because in this example it is unknown which possibility is more probable it will be assumed that Pi3K is with the same probability '+' and '-' (the case where activity of Pi3K is not changed is still included, because '+' and '-' also contain the case that no change occurs at all).

- Hence some additional rules and an additional symbol have to be introduced to the symbols '0', '+', '-' that are already used to describe the calculated states of molecules in the network.

  uk : The activity of the molecule cannot be calculated exactly, therefore it is considered to be 50% '+' and 50% '-'.

  Rule 6 (Calculate 'uk') *'uk' is calculated as value for a molecule using the topology of Transpath if:*

  1. *at least one activating parent is '+' and at least one activating parent is '-'.*

2. *at least one inhibiting parent is '+' and at least one inhibiting parent is '-'.*
3. *at least one activating parent is '+' and at least one inhibiting parent is '+'.*
4. *at least one activating parent is '-' and at least one inhibiting parent is '-'.*

Rule 7 (Probabilities 'uk') *If a molecule receives the value 'uk' then two new networks with the same topology are created, one where the molecule has the value '+' and one where it has the value '-', the newly generated networks share the probability equally between each other (see example described below in text). The previously defined rules 1 – 7 are applied to these generated networks.*

Rule 8 (Probability addition) *The final probabilities of the values of a molecule are calculated by multiplying the value in each network for the molecule and calculating the sum over all networks (see example in text).*

Let us return to the example and apply the new rules. With rule 6.1 Pi3K='uk'. Following rule 7 the network Pi3K+, where Pi3K receives the value '+' and the network Pi3K-, where Pi3K='-' are generated. Both networks have the weight 0.5 and are calculated according to the rules. Pi3K+ leads to the result EpoR='+', Vav='-', Stat5='+', Pi3K='+', Grb-2='0', PIP3='+', SOS='-' and Pi3K- calculates to: EpoR='+', Vav='-', Stat5='+', Pi3K='-', Grb-2='0' ,PIP3='-', SOS='+'. The results obtained by adding the results from Pi3K+ and Pi3K- as described in rule 8 are shown in table 2.1.

To further clarify rules 7 and 8 let all molecules be '0' and the initial conditions be: EpoR='+', Vav='-', Grb-2='+'. Similar to the previous example rule 7 leads 1st to construction of the networks Pi3K+ and Pi3K-, each with the weight 0.5. As SOS='uk' for Pi3K+, Pi3K+ is subdivided into Pi3K+SOS+, where SOS='+' and Pi3K+SOS-, with SOS='-', the weight 0.5 of Pi3K+ is distributed equally between Pi3K+SOS+ and Pi3K+SOS-7, thus each network receives a weight of 0.25. The final result obtained by rule 8 is shown in table 2.2.

The rules and symbolism introduced so far allow to calculate possible networks together with a probability, given some initial conditions and a topology.

However, the rules described until now are only applicable to non cyclic processes and have to be extended to cyclic ones.

Molecule Name	Result in Pi3K+	Result in Pi3K-	Final Probability '+'	'-'	'0'
EpoR	'+'	'+'	1	0	0
Vav	'-'	'-'	0	1	0
Stat5	'+'	'+'	1	0	0
Pi3K	'+'	'-'	0.5	0.5	0
Grb-2	'0'	'0'	0	0	1
PIP3	'+'	'-'	0.5	0.5	0
SOS	'-'	'+'	0.5	0.5	0

Table 2.1: Simple example for rule 8
Pi3K+ is the network, where Pi3K is set to '+' and Pi3K- is the network, where Pi3K is set to '-'. The values for EpoR, Vav, Stat5 and Grb-2 are independent from the value of Pi3K, therefore their value is certain (probability = 1) , for the examined conditions. The values of PIP3 and SOS are influenced by Pi3K, therefore their values are not certain and are influenced by the probabilities for Pi3K+ and Pi3K-, in the example it is assumed, that both networks have the same probability, hence PIP3 and SOS are half '+' and half '-'.

Molecule Name	Result in Pi3K+		Result in Pi3K-	Final Probability '+'	'-'	'0'
	SOS+	SOS-				
EpoR	'+'	'+'	'+'	1	0	0
Vav	'-'	'-'	'-'	0	1	0
Stat5	'+'	'+'	'+'	1	0	0
Pi3K	'+'	'+'	'-'	0.5	0.5	0
Grb-2	'+'	'+'	'+'	1	0	0
PIP3	'+'	'+'	'-'	0.5	0.5	0
SOS	'+'	'-'	'+'	0.75	0.25	0

Table 2.2: More complex example for rule 8
Three networks exist: The network Pi3K+, with a probability of 0.5 , that is further subdivided into SOS+ and SOS-, with a probability 0.25 each and the network Pi3K-, with a probability of 0.5. The final probability for SOS= '+' calculates as follows: (Probability from Pi3K+SOS+) + (Probability from Pi3K-) = 0.25 + 0.5 = 0.75. The other probabilities are calculated analogously.

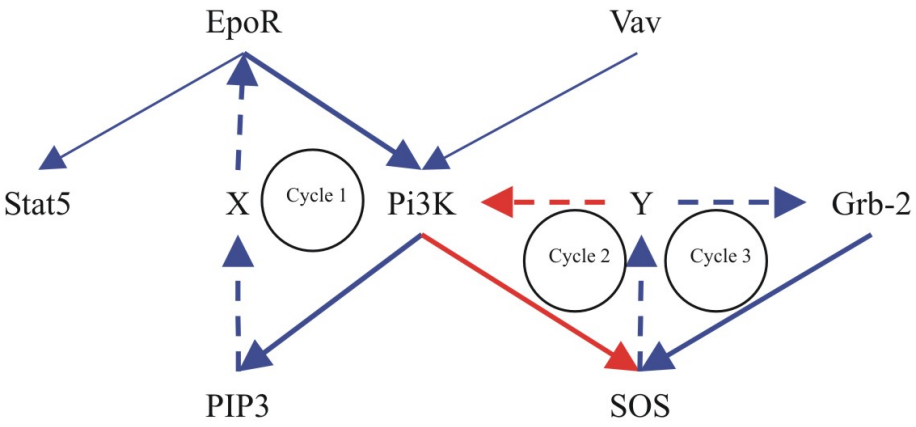


Figure 2.3: Cyclic directed graph
This network is a modification of the network in figure 2.2. 3 cycles were introduced by adding the dotted arrows. Assuming that these minimal cycles play an important role in differentiation, it is logical to search that cycle constellation, that fits best to the given data.

Cyclic graphs

The rules cannot be applied easily to the network shown in figure 2.3, because for elements, which are part of a cycle, rule 7 would have to be used recursively ad infinitum. If positive feedback cycles are the important part for differentiation then it is sufficient to know which cycles exist and which do not, then it is enough to do statistics on the probabilities of cycles.

In the figure there are three positive feedback cycles. Cycle 1 and cycle 2 stabilise the value of Pi3K . Therefore either both cycles will stabilise the value of Pi3K supporting each other or the cycles will compete against each other. In this case one cycle will force its dynamics on Pi3K, while the other one can be neglected. The cycle whose effects are dominated by the other cycle will be called 'inactive' and the dominant cycle will be called 'active'. Using above definitions it can be seen easily, that the activity state of cycle 2 and cycle 3 are depending on each other.

Rule 9 (Allowed Cycles) *1. An 'active' cycle is a cycle consisting of elements with values strongly determined by values of other elements forming this cycle and weak coupling to effects outside that cycle.*

2. *Networks containing cycles that share at least one element are 'allowed' to have all cycles set to 'active' only if all cycles are supporting each other. Contradicting cycles being 'active' at the same time lead to a 'forbidden' network. NW containing contradicting cycles might be 'allowed' if some cycles are 'inactive'.*

Possibility	State of cycle 1	State of cycle 2
allowed	$+\mathit{+}++$	$+--$
forbidden	$+\mathit{+}++$	$-++$
forbidden	$-\mathit{-}--$	$+--$
allowed	$-\mathit{-}--$	$-++$
allowed	$+\mathit{+}++$	inactive
allowed	$-\mathit{-}-+$	inactive
allowed	inactive	$+--$
allowed	inactive	$-++$
allowed	inactive	inactive

Table 2.3: Allowed constellation for cycle 1 and cycle 2
The emphasized symbol represents Pi3K. ++++ for State of cycle 1 means: EpoR= '+', Pi3K= '+', PIP3 = '+' and X= '+', analogously the order for cycle 2 is Pi3K, SOS, Y. Two active cycles are allowed only if the common members (here Pi3K) receive the same values in both cycles. If this is not the case either one cycle is inactive or the constellation is not allowed.

3. *An 'inactive' cycle is a cycle having at least one element that is strongly influenced by dynamics outside of the cycle. These cycles are just topological cycles but not functional because feedback is corrupted.*

Rule 10 (Allowed Networks) *The networks created following rule 7 are valid results if they are 'allowed' by rule 9.*

Rule 11 (Cyclic Probabilities) *Rule 8 is changed slightly to: The final probabilities of the values of a molecule are calculated by multiplying the value in each 'allowed' network for the molecule and calculating the sum over all 'allowed networks'. The sum over all probabilities must be 1, therefore the final results get renormalised.*

Cycle 1 and cycle 2 share the protein Pi3K, thus the both positive feedback cycles can be active only under certain circumstances:

- If cycle 1 is in a state where all its elements have the value '+' then cycle 2 is either in the state$+--$ (Pi3K=+, SOS=-,Y=-) or 'inactive'. If cycle 2 was in the state$-++$, then it would counteract cycle 1. Pi3K would follow the dynamics of the stronger cycle and thus activate it. If both cycles are equally strong then both cycles are 'inactive', because none determines the dynamic of Pi3K. All possible constellations are shown in table 2.3:

- Similar are the relations between cycle 2 and cycle 3, the cycles can only coexist in an activated state in only two constellations, cycle 2 is

in state$-++$ and cycle 3 in state $+++$ or cycle 2 in state $+--$ and cycle 3 in state $---$. The allowed states are analogous to those shown in table 2.3

If one assumes that initially all allowed constellations have the same probability then using the described logic enables one to create an algorithm that tries to find the allowed network that explains most observations and that is self sustaining.

2.3 Computational implementation

To calculate large networks according to rules 1–11 a computer programme was written. There are multiple ways how this can be done, therefore the main programme steps are described in this chapter.

Step 1 (Set Default:) *Initially all elements of the network have the value '0' and thus do not change the network (rule 1)(see figures 2.4 a, 2.5a).*

Step 2 (Initialise with known values:) *If gene expression is 'IncrExpr' or 'DecrExpr' for a certain gene then the value for the gene and the protein it is encoding is set to '+' or '-' (rules 2, 4) in accordance with the measurement (see also figures 2.4b, 2.5b). The programme stores values of this network in a vector called (**PreviousNet**).*

Step 3 (Calculate consequences:) *The elements having the values '+' or '-' effect their children by inhibiting and activating relationships (rules 3, 5, 6). The update is done by executing the rules derived from the arcs in the directed graph in the order they are written in a table. The rules used in the programme are based on rules 3, 5, 6. However, rules 3, 5, 6 had to be modified in such a way that their previous value is taken into consideration. This is important because it ensures, that a cycle that was chosen to be 'inactive' cannot become 'active' by chance.*

Thus the rules are modified as follows (examples in figures 2.4b-g, 2.5b,c):

1. *Parents having the values 'unchanged' do not change the value of their children (rule 1).*
2. *If the parent has the value '+' and the relation is activating then:*
   (a) *The child receives the value '+' if the previous value of the child was '0' or '+'.*
   (b) *The child receives the value 'uk' if its previous value was '-' or 'uk'.*
3. *If the parent has the value '+' and the relation is inhibiting then:*

   (a) *The child receives the value '-1' if, the previous value of the child was '0' or '-'.*

   (b) *The child receives the value 'uk' if its previous value was '+' or 'uk'.*

4. *If the parent has the value '-' and the relation is activating then:*

   (a) *The child receives the value '-' if the previous value of the child was '0' or '-'.*

   (b) *The child receives the value 'uk' if its previous value was '+' or 'uk'.*

5. *If the parent has the value '-' and the relation is inhibiting then:*

   (a) *The child receives the value '+' if the previous value of the child was '0' or '+'.*

   (b) *The child receives the value 'uk' if its previous value was '-' or 'uk'.*

Execution of these rules leads to a network called ***ActualNet****.*

Step 4 (Assign random values to 'uk')

After the entire list has been executed the algorithm assigns other values to the elements having the value 'uk' in ***ActualNet****:*

1. *If the value of the element was '0' in* ***PreviousNet*** *then the element gets by chance the value '+' or '-'. The probability is the same for both values (figures 2.5d, 2.6a).*

2. *If the value of the element was '+' or '-' in* ***PreviousNet*** *then the algorithm checks if the element has at least one parent supporting its previous value. If such a parent exists, then the element gets the same value as in* ***PreviousNet*** *(figure 2.5d), otherwise either '+' or '-' gets assigned randomly to the element. The algorithm contains a factor that determines the chances. This makes it possible to favour the previous value. I kept in my calculations the probabilities for '+' and '-' equal.*

By assigning values to 'uk' elements ***ActualNet*** *is changed. The programme remembers to which molecules it assigned random values and also the assigned values.*

Step 5 (Repeat till no change or stable oscillations occur:) *If* ***ActualNet*** *and* ***PreviousNet*** *are not the same and if the system is not showing constant oscillations (determined by comparing* ***ActualNet*** *with the* ***ActualNet*** *received ten iterations before) then* ***PreviousNet*** *is set to equal* ***ActualNet*** *and the algorithm returns to step 3 (see also figure 2.7).*

Step 6 (Scoring against data:) *The resulting network is scored against the measured values. The more values in the network equal the measured values the better the network (figure 2.8).*

Step 7 (Repetition:) *The algorithm is repeated multiple times from step 1 to find the networks that get the highest score at step 6.*

Scoring for a single measurement point

In step 6 the results, obtained by calculations performed by initialising the network with the measured data as described under step 2, are again compared with this data. Each value of an expressed protein or gene, that corresponds to rules 2,4 (step 2) gives one point. The points are added to receive the score for the complete network. In figure 2.8 an example is given, under the assumption that measurements show EpoR= '+' and Vav= '-'.

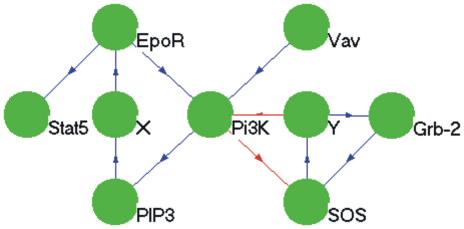
2.3.1 Modifications to the algorithm enabling it to cope with time series

The above algorithm can be used with single measurements but as time series are available it is useful to change it in such a way that it can compute time series. For this purpose let us consider the network shown in figure 2.9 and assume that the observations made with gene chips are those shown in table 2.4. In addition to the information in the table lets assume that it is also known that EpoR= '+'.

As some genes are not expressed initially, they and the proteins they are producing are removed from the network (see figure 2.9b). The changes in expression shown in the second column of table 2.4 are the difference in gene expression between the second and the first time point so, they occur after EpoR exerted its effects. Therefore, the network is initialised solely with EpoR= '+' (see figure 2.10.a) and the simulation is done as described in the previous chapter (step 1–7).

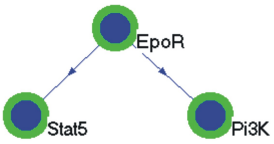
In bigger networks and especially if the experimentally measured time points have a large distance to each other, gene effects might be involved in the process. Therefore, after the 1st stable network has been calculated all values are set to '0' and the network is then reinitialised with EpoR= '+' and the genes having the values calculated in the previous simulation (see figure 2.10.c). This recursive procedure is repeated several times. The results for human cells show that the previous calculation restricts solution space of the next simulation. Some states can be obtained only by passing multiple turns of calculations described in the previous section.

After multiple calculations for one time point, the measured changes in expression are introduced into the calculation (figure 2.11.a). This means that, after setting all values to '0' the network is initialised with measured

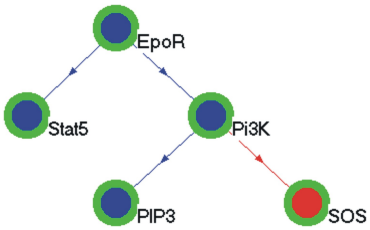


(a) Initially all molecules have the value '0' (step 1).

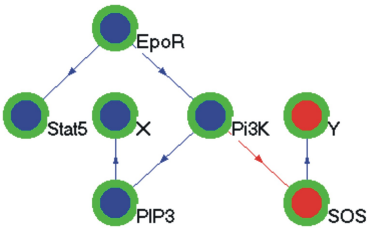
(b) Let EpoR be '+'. As '0' elements of the network do not change their children and because '0' is the default value, those elements are not shown.



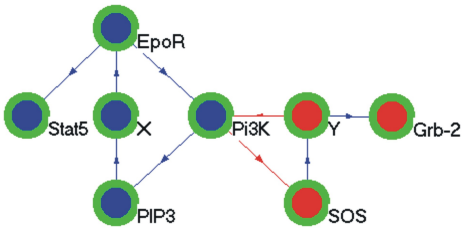
(c) EpoR influences at first the activity of Stat5 and afterwards the activity of Pi3K (step 3.2.a).



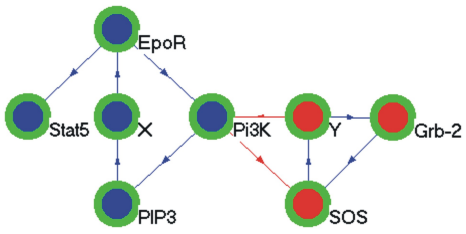
(d) The results for the next time step are calculated: Pi3K sets PIP3 to '+' (step 3.2.a)) and SOS to '-' (step 3.3.a).



(e) Normal calculation of consequences with steps 3.2.a, 3.3.a.

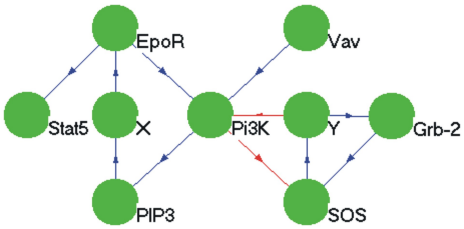


(f) X sets value of EpoR to '+' in a positive feedback loop (step 3.2.a)). Because X is supporting the value that EpoR already has, it is not necessary to change the network.



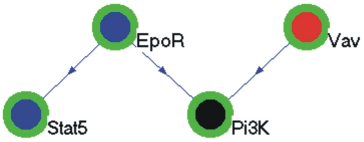
(g) This is the network state resulting from EpoR= '+'. Stat5, X, Pi3K, PIP3 are also '+' and Y, Grb-2 and SOS get '-'.

Figure 2.4: Time Series for Non-Contradicting reactions
Colours have the following meaning: green= '0', blue= '+', red= '-'. Core colour symbolises current value, border colour symbolises previous value of node



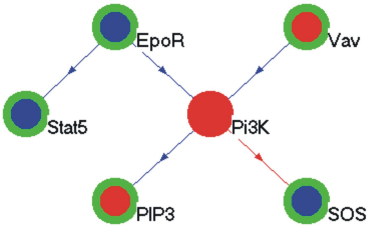
(a) Initially all molecules have the value '0' (step 1).

(b) Let us assume that observation shows that EpoR= '+' and Vav= '-'. How do the values of the network change ?

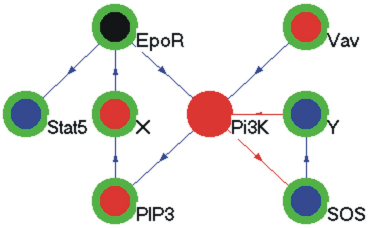


(c) The algorithm recognizes that the inputs for Pi3K are contradicting each other, therefore Pi3K= 'uk' (steps 3.2.a , 3.4.b)

(d) As the algorithm cannot calculate the exact value it assigns a random value to Pi3K (in this case '-') and stores this decision in a variable (step 4.1) (this is indicated by the red border around Pi3K).

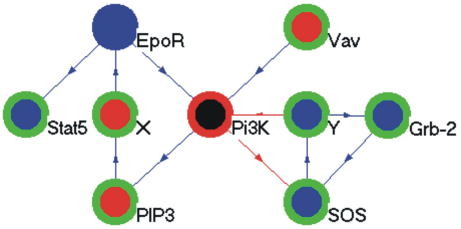


(e) The algorithm calculates the consequences of that assignment (step 3).

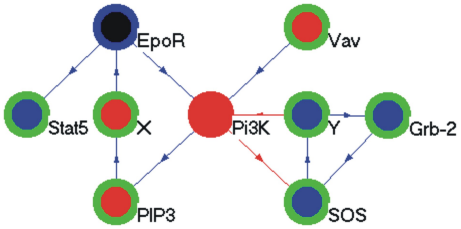


(f) The algorithm realizes that EpoR is '+' but should be '-', because of X, therefore similar to d.), EpoR= 'uk' (step 3.4.b.) in step 4.1 a random value will be assigned to EpoR. This time series is continued in figure 2.6

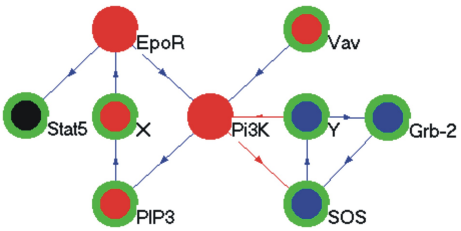
Figure 2.5: Time series for contradicting reactions (Part 1)
Symbols are the same as in the previous figure but 'uk'= black has been added. Colours have the following meaning: green= '0', blue= '+', red= '-'. Core colour symbolises current value, border colour symbolises previous value of node



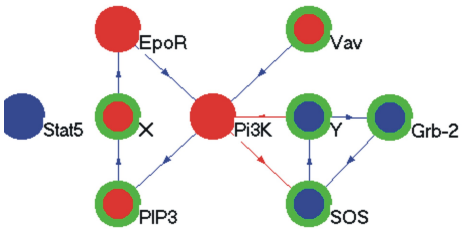
(a) Continuation of figure 2.5. The algorithm assigns by chance the value '+' to EpoR and remembers that it decided that way, as indicated by the blue border around EpoR (step 4.1). As a result, there is again a conflict with the value of Pi3K.



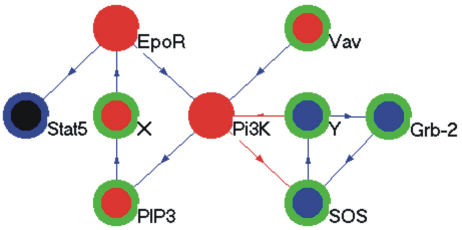
(b) The algorithm knows that the value of Pi3K is not consistent with the entire network but as it also knows that it has already assigned the value '-' Pi3K by chance (see figure 2.5 d) it first checks if there is at least one parent supporting the previous, randomly assigned value. As Vav supports the value '-' for Pi3K the algorithm assigns again '-' to Pi3K (step 4.2) and recognizes that there is again a conflict with EpoR.



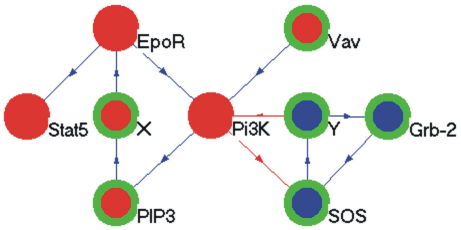
(c) Analogously to b) the algorithm checks if EpoR has at least one parent supporting the value '+'. Because this is not the case it assigns the value '-' to EpoR and remembers this (step 4.2). As a result of the changed value for EpoR there is now a conflict with the previous value of Stat5.



(d) The algorithm assigns a random value to Stat5 and remembers which value it has assigned.

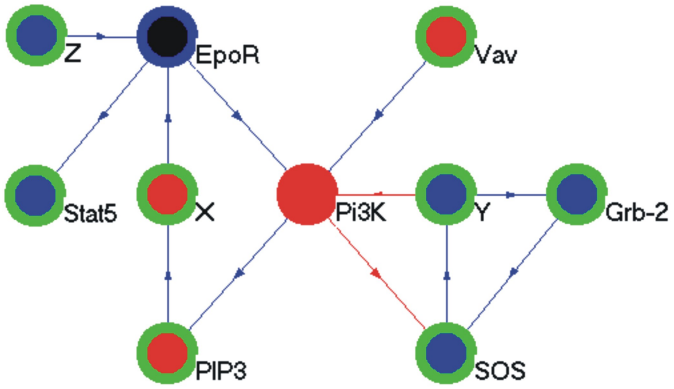


(e) The value of Stat5 is not consistent. Now the algorithm checks, if Stat5 has at least one parent supporting its previous value and cannot find one ...

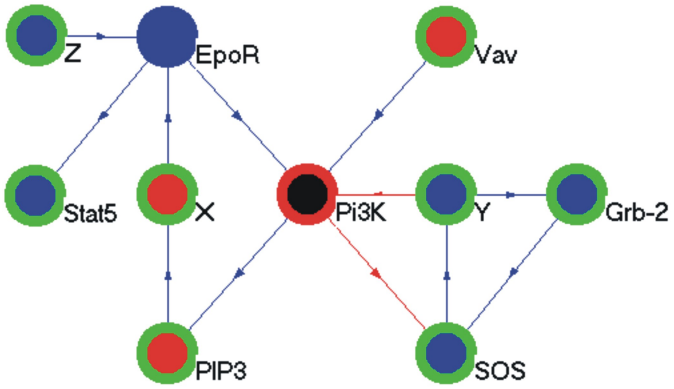


(f) ... therefore it assigns the opposite value to Stat5 and remembers this. This time the network is consistent and the algorithm gives this as a result.

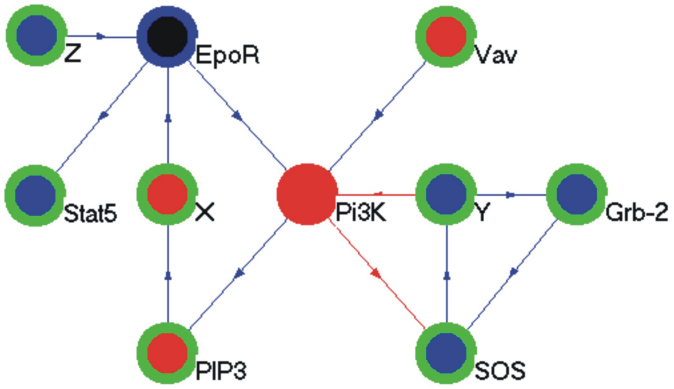
Figure 2.6: Time series for contradicting reactions (Part 2)
This example shows how the application of the algorithm handles contradicting reactions by making cycles 'inactive'. Pi3K is a common element between Vav and the cycle EpoR, Pi3K, PIP3, X, EpoR. In b the algorithm by chance sets the cycle to 'inactive', by giving effect of Vav a higher weight. In c it realises that EpoR= '+' is inconsistent if the cycle is 'inactive', therefore the value of EpoR gets changed.



(a) This is a modified form of figure 2.6 b. EpoR has a parent supporting its previous value.



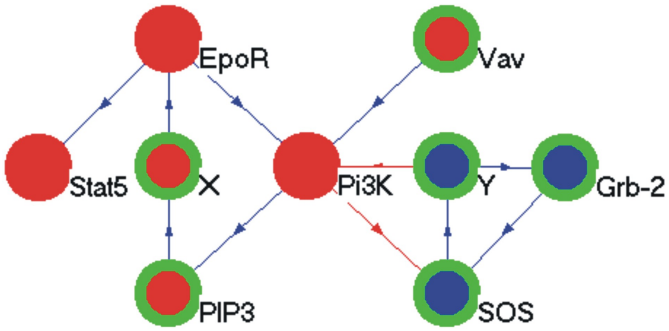
(b) As the previous value of EpoR is supported it receives again the same value. Consequently, the program detects a conflict with the value of Pi3K. As Pi3K's previous value is supported by Vav...



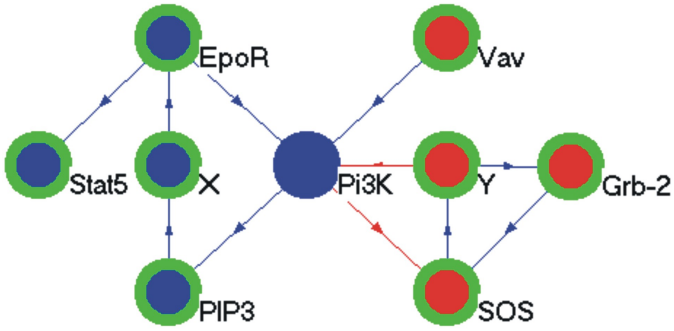
(c) ... the algorithm keeps the previous value of Pi3K and produces again a conflict with the value of EpoR. Here the algorithm realises that the network in subfigure a) and the network in subfigure c) are the same and thus stops simulation. The result is an oscillating network.

Figure 2.7: Time series for an oscillating System

(a) If experiments show that EpoR= '+' and Vav= '-', then the programme calculates two solutions



(b) As shown in previous figures, this is one possible solution. Scoring this solution against the measured results would give 1 point for Vav, because it is '-' in the simulation and in the observation and 0 points for EpoR because network result and observation contradict each other.



(c) This is the 2nd possible solution. Here EpoR as well as Vav have the same values in the simulation and the experiment. Therefore this solution scores 2 points and thus is better than the previous one.

Figure 2.8: Scoring resulting networks

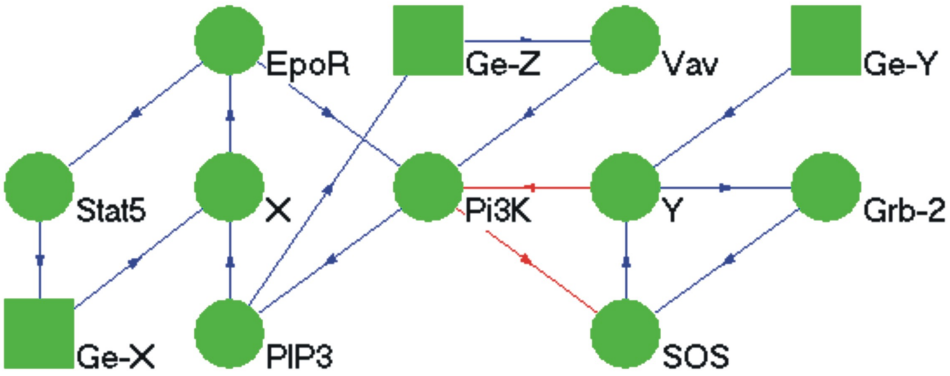
values (in the example, those from table 2.4). The resulting network is reinitialised and recalculated as described in the previous paragraph (figure 2.11).

After calculating several results using the algorithm described in the previous section the unexpressed genes and their proteins are removed from the network and everything is repeated as described above, till all measurement points have been inserted.

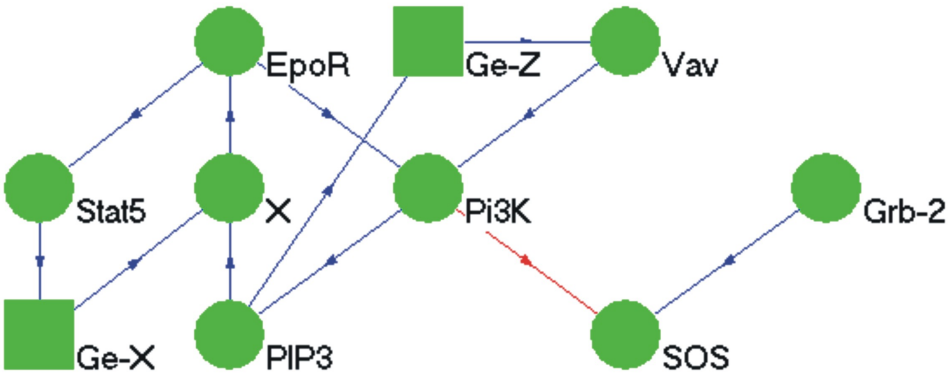
Throughout the simulations the programme calculates how well computed results agree with the experiments and how many elements of the network change their value during the simulation. Finally, after multiple runs were performed that way, the best results, being those with a high similarity to the experimental data and a low number of changes are stored for further calculations. An example for three runs is shown in table 2.5.

Time Point	Gene Name	Observation	
		Expression Status	Difference in Expression
1	Ge-X	'PRESENT'	'IncrExpr'
	Ge-Y	'ABSENT'	No Change in Expression
	Ge-Z	'PRESENT'	'DecrExpr'
2	Ge-X	'PRESENT'	'IncrExpr'
	Ge-Y	'ABSENT'	No Change in Expression
	Ge-Z	'ABSENT'	No Change in Expression
3	Ge-X	PRESENT	–
	Ge-Y	ABSENT	–
	Ge-Z	PRESENT	–

Table 2.4: Example expression data



(a) Let this be a modified network, consisting of Genes Ge-X, Ge-Y, Ge-Z.

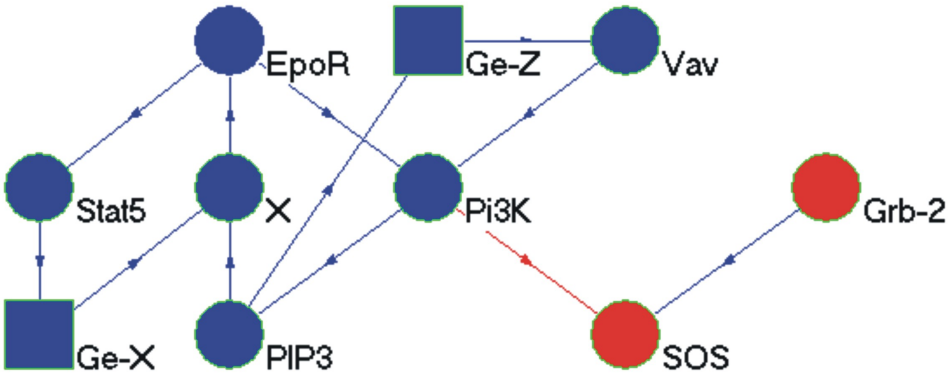


(b) Because Ge-Y is ABSENT (see 2.4) it is removed from the network.

Figure 2.9: Network with genes

Circles represent molecules, squares represent genes. Molecules and genes are allowed to take the following values '+', '-', 'uk', '0'. In addition to afore mentioned values genes can also be 'PRESENT' or 'ABSENT'. 'ABSENT' genes and the proteins they are encoding are removed from network (see b).

(a) EpoR is '+' as mentioned in the text. The other available information is the difference in gene expression between 2nd, 1st and 3rd, 2nd experimental time point (see table 2.4). It is assumed, that signalling effects are faster than effects originating from change in expression and therefore the network is just initialised with the information on the signalling level.

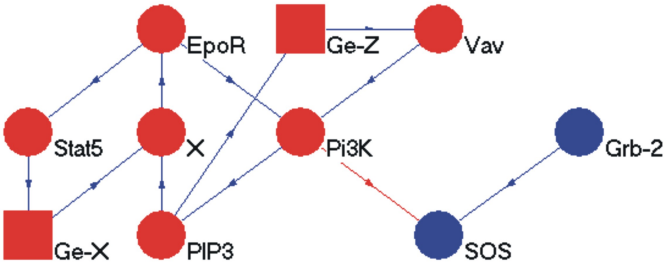


(b) Then this is the only possible solution.

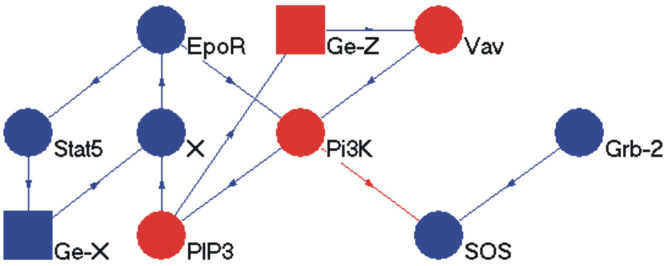
(c) The network gets reinitialised with the calculated values. In this case the solution is still the same as for a). The idea behind this recalculation is that initially the fast signalling events play an important role and regulate the genes but in the long run the dynamic of the genes starts to dominate. These calculations show the effects of EpoR. Assuming that the model does not cover all relevant events, the network is again recalculated but this time it is reinitialised by the measured difference in expression for t2,t1 (see table 2.4) This is shown in figure 2.11

Figure 2.10: Network with genes t1

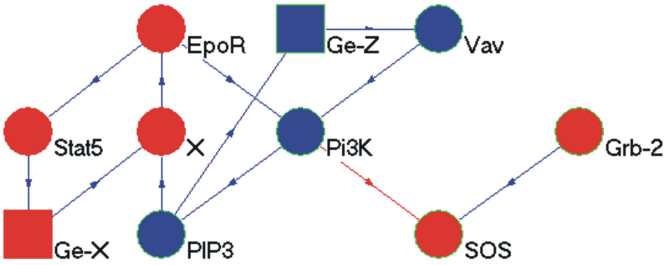
(a) Table 2.4 shows that Ge-X= 'IncrExpr' and Ge-Z= 'DecrExpr'. As it is known that EpoR= '+' the network has to be initialised as shown in the figure. This network can have multiple solutions. One solution is the one shown in figure 2.9 a. The other three are shown here.



(b) This is the solution obtained if Ge-Z is the dominant factor.



(c) This is the solution obtained if all factors have the same strength.



(d) As the algorithm decides randomly, this is also a possible solution.

Figure 2.11: Network with genes t2

(a) From figure 2.11 different reinitialisation possibilities can be derived. The 1st one is shown here. This solution will produce for every timepoint the solution shown in figure 2.9 b) as result.

(b) This is the reinitialisation resulting from figure 2.11 b). As EpoR is known to have an increased activity and because it is not a gene it is set to increased although the network computed decreased as result. The possible outcomes of this network are all four networks.

(c) This is the reinitialisation resulting from figure 2.11 c). Here also all four networks are possible results.

(d) This is the reinitialisation resulting from figure 2.11 d). Here also all four networks are possible results.

Figure 2.12: Initialisation possibilities

Run	Meas. Time	Sim. Time	Value for Elements in stable state												∑Score	∑Changes
			EpoR	Ge-Z	Vav	Ge-Y	Stat5	X	Pi3K	Y	Grb-2	Ge-X	PIP3	SOS		
1	1	1	+	+	+	0	+	+	+	0	-	+	+	-	1	0
		2	+	+	+	0	+	+	+	0	-	+	+	-	2	0
		3	+	+	+	0	+	+	+	0	-	+	+	-	3	0
	1.5	4	+	+	+	0	+	+	+	0	-	+	+	-	6	0
		5	+	+	+	0	+	+	+	0	-	+	+	-	9	0
		6	+	+	+	0	+	+	+	0	-	+	+	-	12	0
	2	7	+	0	0	0	+	+	+	0	-	+	+	-	15	2
		8	+	0	0	0	+	+	+	0	-	+	+	-	18	2
		9	+	0	0	0	+	+	+	0	-	+	+	-	21	2
2	1	1	+	+	+	0	+	+	+	0	-	+	+	-	1	0
		2	+	+	+	0	+	+	+	0	-	+	+	-	2	0
		3	+	+	+	0	+	+	+	0	-	+	+	-	3	0
	1.5	4	+	—	-	0	+	+	-	0	+	+	-	+	7	6
		5	+	—	-	0	+	+	-	0	+	+	-	+	11	6
		6	+	—	-	0	+	+	-	0	+	+	-	+	15	6
	2	7	+	0	0	0	+	+	+	0	-	+	+	-	18	12
		8	+	0	0	0	+	+	+	0	-	+	+	-	21	12
		9	+	0	0	0	+	+	+	0	-	+	+	-	24	12
3	1	1	+	+	+	0	+	+	+	0	-	+	+	-	1	0
		2	+	+	+	0	+	+	+	0	-	+	+	-	2	0
		3	+	+	+	0	+	+	+	0	-	+	+	-	3	0
	1.5	4	+	—	-	0	+	+	-	0	+	+	-	+	7	6
		5	-	+	+	0	-	-	-	0	-	-	+	-	7	16
		6	+	+	+	0	+	+	+	0	-	+	+	-	10	20
	2	7	+	0	0	0	+	+	+	0	-	+	+	-	13	22
		8	+	0	0	0	+	+	+	0	-	+	+	-	16	22
		9	+	0	0	0	+	+	+	0	-	+	+	-	19	22

Table 2.5: Example run

For Sim. Time 1 the network was initialised with EpoR= '+'. In Sim. Time 2–3 initialisation was done with EpoR= '+', GE-Z= '+' and Ge-X= '+' (see figure 2.4). During Sim. Time 1–6, which corresponds to Meas. Time 1 and 1.5 Ge-Y and Y are removed from the network (here denoted with '0'), because according to table 2.4 Gene Ge-Y is 'ABSENT' during the 1st time point. For Meas. Time 1.5, corresponding to Sim. Time 4–6, the simulation is initialised with EpoR= '+' and as follows from Difference in Expression for Time Point 1 in table 2.4 Ge-X= '+' and GE-Z= '-' (see also figures 2.11 and 2.12). Emphasised values are calculated values agreeing with expected values from table 2.4 for denoted time point. Each hit gives one point. The points are accumulated, because each Sim. Time point evolves from the previous one.

Further analysis of obtained results

From table 2.5 one can see that the 2nd run with 24 points delivers the network matching best to the data shown in table 2.4. By contrast, the worst network is the last one. It has the lowest score (just 19) and a high number of changes in protein values. The 1st run has a score of 21 and just 2 changes. Because of this the network obtained from the 1st and 2nd run are chosen as good networks for further analysis.

The allowed networks are treated as similar good solutions and the algorithm tries to find common parts in the solution. The calculations are performed for the entire time (simulation time 1–9) and for the 3 single intervals (e.g. simulation time 1–3; 4–6; 7–8).

The stability of the values is calculated as the average value of the observed molecule over the observed time. For example, the value stability of Vav during simulation time 1–9 is calculated as

$$1 + 1 + 1 + 1 + 1 + 1 + 0 + 0 + 0 = 6$$

from the 1st run and

$$6 + 1 + 1 + 1 - 1 - 1 - 1 + 0 + 0 + 0 = 6$$

from the 1st run added to the 2nd run, leading to the result:

$$6/18 = 0.3333$$

As the value 0.333 is low, the calculations for the value of Vav are not considered very reliable. By contrast, for EpoR this value is 1, therefore, it is very likely that EpoR has an increased activity over the entire calculated time. Observing Vav only for the 1st three simulated time points leads to the reliable solution that Vav= '+' (it calculates as $\frac{3+3}{6} = 1$). These calculations are performed for every element for all relevant time periods. All the results for the example can be found in table 2.6.

Additionally, the correlation coefficients between all molecules and relevant time periods are calculated. These are used to identify relevant reactions in the network. For example, in the 2nd run the value of Grb-2 determines the value of Pi3K, therefore, the programme would calculate a lower importance for the interaction between EpoR and Pi3K: Sometimes, as is it the case here for all entries, it is not possible to calculate a correlation coefficient, because the values do not change, in these cases the coefficient for the entire measurement (in this example 1–9) has been taken, if this is not possible like in this example for EpoR, a simple factor has been calculated. Calculation is exemplified for EpoR and Pi3K for the entries in table 2.5. Table 2.7 contains all correlation coefficients for the example.

The programme checks if correlated elements are linked in the graph, if this is the case it checks if the link has the right properties (e.g. elements

with a positive correlation coefficient should be linked by an activating reaction, negatively correlated elements should have an inhibiting link). If this is also the case then the programme assigns the correlation coefficient as a strength to the link,an example is shown in figure 2.13.

$$\frac{1+1+1+1+1+1+1+1+1+1+1+1-1-1-1+1+1+1}{18} = 0.6666$$

Time Period	EpoR	Ge-Z	Vav	Ge-Y	Stat5	X	Pi3K	Y	Grb-2	Ge-X	PIP3	SOS
1–9	1	0.33	0.33	0	1	1	0.66	0	-0.66	1	0.66	-0.66
1–3	1	1	1	0	1	1	1	0	-1	1	1	-1
4–6	1	0	0	0	1	1	0	0	0	1	0	0
7–9	1	0	0	0	1	1	1	0	-1	1	1	-1

Table 2.6: Stability

This table shows the stability values derived from table 2.5. In table 2.5 the 3rd run has a low score and high change and therefore is omitted. Run 1 and 2 have high scores and low changes and therefore are used to calculate stability. Stability is calculated, as the average value of the molecule over the observed time, i.e. stability of Ge-Z for the time period 1–9 is given as: $\frac{[(1+1+1)+(1+1+1)+(0+0+0)]+[(1+1+1)+(-1-1-1)+(0+0+0)]}{18} = \frac{6}{18} = 0.33$

	EpoR	Ge-Z	Vav	Ge-Y	Stat5	X	Pi3K	Y	Grb-2	Ge-X	PIP3	SOS
EpoR	1	0.50	0.50	-	1	1	0.67	-	-0.67	1	0.67	-0.67
Ge-Z	0.50	**1**	**1**	-	0.50	0.50	**1**	-	**-1**	0.50	**1**	**-1**
Vav	0.50	**1**	**1**	-	0.50	0.50	**1**	-	**-1**	0.50	**1**	**-1**
Ge-Y	-	-	-	-	-	-	-	-	-	-	-	-
Stat5	1	0.50	0.50	-	1	1	0.67	-	-0.67	1	0.67	-0.67
X	1	0.50	0.50	-	1	1	0.67	-	-0.67	1	0.67	-0.67
Pi3K	0.67	**1**	**1**	-	0.67	0.67	**1**	-	**-1**	0,67	**1**	**-1**
Y	-	-	-	-	-	-	-	-	-	-	-	-
Grb-2	-0.67	**-1**	**-1**	-	-0,67	-0,67	**-1**	-	**1**	-0,67	**-1**	**1**
Ge-X	1	0.50	0.50	-	1	1	0,67	-	-0,67	1	0,67	-0,67
PIP3	0,67	**1**	**1**	-	0,67	0,67	**1**	-	**-1**	0,67	**1**	**-1**
SOS	-0,67	**-1**	**-1**	-	-0,67	-0,67	**-1**	-	**1**	-0,67	**-1**	**1**

Table 2.7: Correlation coefficients

The correlation coefficients are calculated, using the data from table 2.5, substituting '+' with 1 and '-' with -1 and applying the usual formulae for the correlation coefficient. If the formulae was not applicable then alternative methods described in the text were applied. The correlation coefficients are compared with reactions in the database do identify, dominant interactions between elemets.

2.3.2 Sensitivity analysis on a discrete network

A simplified algorithm was used to estimate the sensitivity of the network based on change in certain molecules. The main change to the algorithm described in section 2.3 is, that elements of the network which received the value 'uk' keep this value.

The estimation is based on the following ideas:

- Signal transduction downstream of EpoR is a more linear process, which is constrained by inhibited elements.
- Only elements which have a direct inhibiting reaction can be inhibited completely.
- As the inhibition probability is not known it is chosen randomly, by this process many possible networks are generated.
- The likelihood that an element is important for the real network is biggest for those elements that are important also for the set of possible networks.

These assumptions are to some degree in contrast to the cyclic assumptions made for the previous sections.

The algorithm is as follows (figures 2.14,reffig:SensitivityAnalysis2 show how the algorithm is working):

1. Identify genes being 'ABSENT' and remove those (and proteins as described in section 2.2.3) from the network.
2. Make a list, called **diakoptic list** of elements, which have at least one inhibiting parent.
3. Define an inhibition probability **Prob.** for molecules being on the **diakoptic list**. Start with 0% inhibition probability and increase it in 10% steps to 100%.
4. Set the 1st element in the **diakoptic list** to inhibited and inhibit randomly elements on the **diakoptic list** with regard to **Prob.**. Inhibited elements do not change the values of their children.
5. Calculate the resulting values using the rules from section 2.3 with the specific modifications.
6. Remember the values for the resulting network.
7. Perform point 5–6, but this time let the element, that was set to inhibited in point 4, influence its children.
8. Repeat point 4 to 6 1000 times.

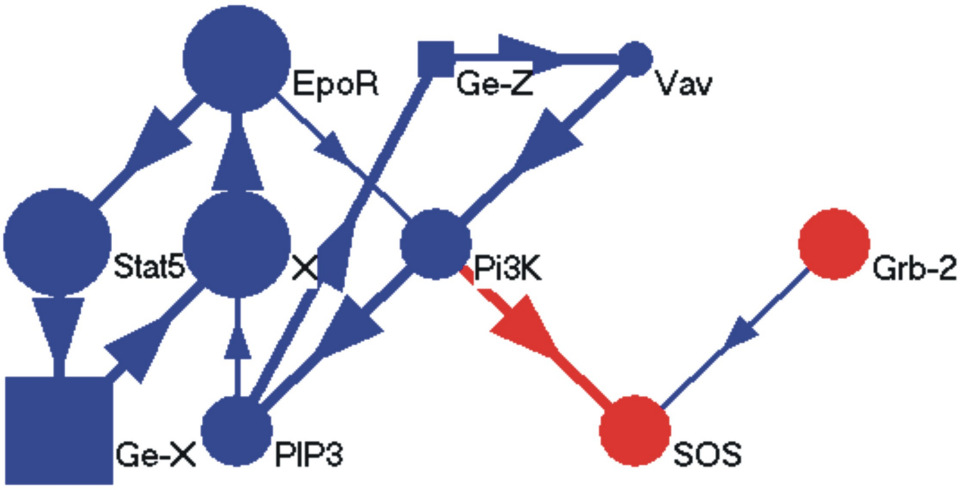


Figure 2.13: Example of calculated data shown in graph
The thickness of the arrows indicates the correlation coefficient between the two elements connected by the arrow, as long as the reaction favours that correlation (e.g. if the correlation coefficient is negative and the arc is an activating arc, then the arrows get a small default value 0,1). The size of the nodes indicates in how many 'good' networks a node receives the specified value.

9. Calculate the differences (see below) from the stored values and save them in a file.

10. Go to point 4 and repeat all with the 2nd, 3rd etc. element of **diakoptic list** till all elements were set once to inhibited and once to not inhibited.

11. Go to point 3 and increase inhibition probability by 10%.

The following values were calculated from the saved data for each element in the **diakoptic list**:

Difference in Network Size: This difference is the difference in the number of molecules still attached to EpoR for the network, where the tested molecule is inhibited and the size of the network, where this molecule is not inhibited. This value is than averaged over all 1000 simulations (see step 8).

Difference in Network Behaviour: This distance is defined as follows: If the observed molecule has another value than 'uk' then the difference is 0. If the value is 'uk' then the algorithm creates two new networks, one where the observed molecule has the value '+' and one network where it is '-'. Both networks are calculated as described previously.

The results of each molecule for both networks are compared with each other to calculate this difference. If a molecule is '+' in one calculation and '-' in the other then the difference of this molecule is one. If the value of the molecule in one simulation is 'uk' and either '+' or '-' in the other then the difference is 0.5. In all other cases the distance between the molecules is 0. The distance between two networks, is the sum of the distances of their molecules. This value is also averaged over all 1000 simulations.

Summarising the results into a single factor, to rank them

The results of the analysis were summarised by assigning a single numerical value **RF** to each molecule in the **diakoptic list**, as a measure of its importance.

$$RF_i = \sum_{Prob=0\%}^{100\%} \frac{MeanAll_{Prob} - Mean_{i,Prob}}{StdAll_{Prob}}$$

*RF_i: Numerical Value **RF** for molecule* i *in **diakoptic list***
*$MeanAll_{Prob}$: Mean Value for the Mean Values of all molecules in **diakoptic list**, when the probability of inhibition is* Prob.
$Mean_{i,Prob}$: Mean Value for i *if inhibition probability is* Prob
*$StdAll_{Prob}$: Standard Deviation for the mean values of all elements in **diakoptic list**, for inhibition probability* Prob.

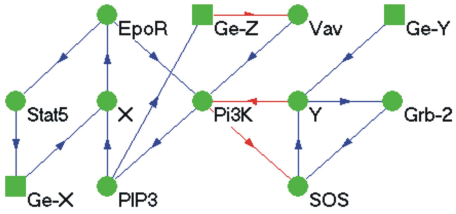
Explanation of results for example

In the example the network shown in figure 2.14a was analysed by applying the algorithm for sensitivity analysis on it. If expression data was available, then the network analysed would be changed by removing unexpressed genes and their corresponding proteins as described in section 2.2.3. For this example it will be assumed, that all proteins are 'PRESENT'.

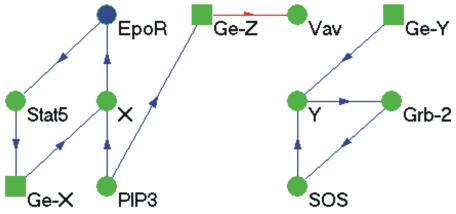
As can be seen easily from figure 2.14a, Vav, Pi3K and SOS will go into the diakoptic list (step 2). The elements in this list are removed from the network with a probability *Prob*, hence $2^3 = 8$ different topologies are possible. The probability of each topology is depending on the value assigned to *Prob* and the examined molecule.

Each single calculation of 'Difference in Network Size' and 'Difference in Network Behaviour' requires two networks, one where the treated molecule is in the network and one where the molecule is removed, the two networks are compared and this procedure is repeated several times.

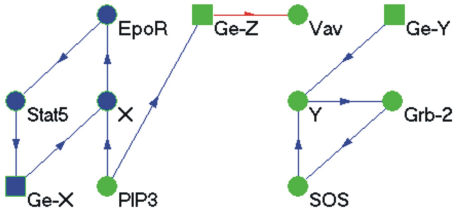
For the example it is possible to calculate the exact probabilities. The results are shown in table 2.8. The algorithm calculates the final values of 'Difference in Network Size' and 'Difference in Network Behaviour' by building the sum of each calculated value for each network and multiplying it



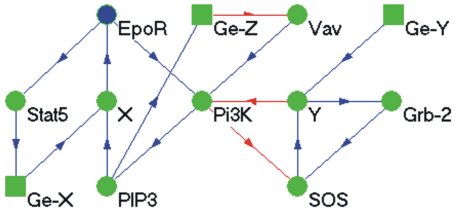
(a) This is the given network. For illustration purposes the reaction between Ge-Z and Vav has been changed from an activating to an inhibiting reaction. The algorithm puts the elements Pi3K, SOS and Vav into the **diakoptic list** (step 2 of algorithm), because each has at least one inhibiting parent.



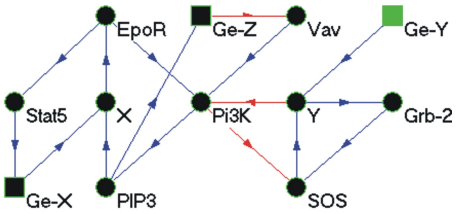
(b) Pi3K is the 1st element in the **diakoptic list** so it is removed 1st (step 4). The elements SOS and Vav are removed with probability **Prob** (also step 4). Here, by chance, both elements are kept in the network.



(c) The algorithm calculates the resulting network. EpoR effects only Stat5, Ge-X and X without Pi3k, all other elements remain '0' (step 5). The network remembers that EpoR influences 3 molecules (step 6).

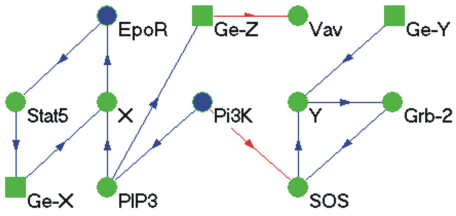


(d) Pi3K is kept in the network and the previous calculations are repeated (step 7).

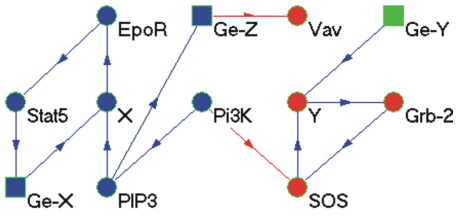


(e) If Pi3K is kept in the network this is the result. EpoR influences now 10 molecules and all of them have the value 'uk' (black). **Difference in Network Size** is calculated by building the difference between the elements influenced in c.) and this network, hence $10-3=7$. This value is remembered by the programme. Because Pi3k, the examined molecule, has also the value 'uk', two new networks are created, one, where Pi3K is forced to be '+' and one where it is forced to be '-'. These networks are shown in figure 2.15.

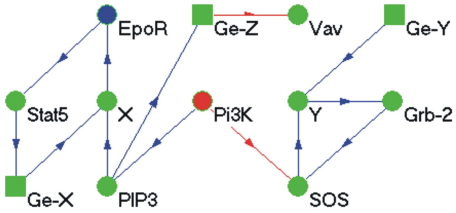
Figure 2.14: Sensitivity analysis (Part 1)



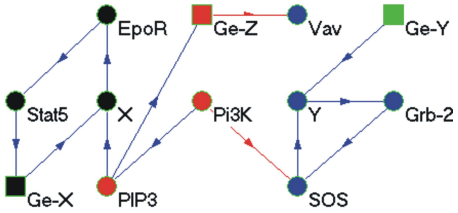
(a) Pi3K is set to be '+'. To ensure that it does not change its value during the simulation its upstream reactions are removed from the network.



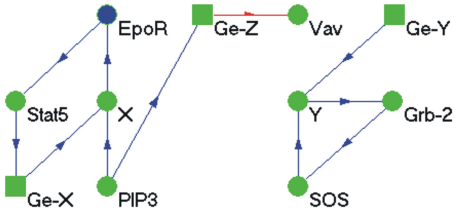
(b) This is the result, if Pi3K= '+'.



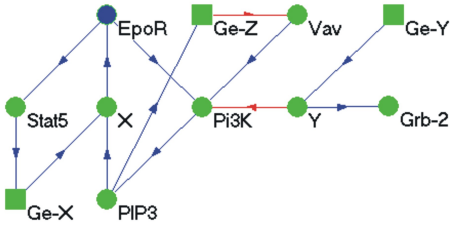
(c) In this network Pi3K is forced to be '-' and it cannot change its value, because all upstream reactions are removed.



(d) This is the result if Pi3K= '-'. The programme calculates the **Difference in Network Behaviour** from this network and network b) of this figure series. The result is: $0.5 + 0.5 + 0.5 + 0.5 = 2$ from EpoR, Stat5, Ge-X and X $1+1+1+1+1+1+1 = 7$ from PIP3, Ge-Z, Vav, Pi3K, SOS, Y, Grb-2. Thus the result is $2 + 7 = 9$.



(e) One Run is completed and the programme returns to point 4, removes Pi3K, and randomly some other elements from the **diakoptic list** and then repeats all calculations. A network, where also SOS is removed but Vav is kept is shown here. This procedure of creating new networks for Pi3K is repeated 1000 times.



(f) After 1000 networks for Pi3K have been calculated, the programme averages over all **Differences in Network Size** and **Differences in Network Behaviour**. Then it repeats the entire procedure for the next element in the **diakoptic list**, in this example it is SOS. If all elements in the list are treated this way the programme increases the inhibition probability **Prob** by 10% and repeats everything again.

Figure 2.15: Sensitivity analysis (Part 2)

Examined molecule	Inhibited secondary element from **diakoptic list**			Diff. in N. Size	Diff. in N. Behaviour	Probability of Network
	Pi3K	SOS	Vav			
Pi3K		-	-	3	0	$Prob^2$
		+	-	6	0	$(1-Prob)*Prob$
		-	+	4	6	$Prob*(1-Prob)$
		+	+	7	9	$(1-Prob)^2$
SOS	-		-	0	0	$Prob^2$
	+		-	3	0	$(1-Prob)*Prob$
	-		+	0	0	$Prob*(1-Prob)$
	+		+	3	3	$(1-Prob)^2$
Vav	-	-		0	0	$Prob^2$
	+	-		1	4	$(1-Prob)*Prob$
	-	+		0	0	$Prob*(1-Prob)$
	+	+		1	5.5	$(1-Prob)^2$

Table 2.8: Explanation for results in example
This table shows Probabilities of the possible networks for each examined molecule, their 'Difference in Network Size' and 'Difference in Network Behaviour'. The tables shows, that the Probability of a network shifts to the 1st network for high Prob-*Values and to the 4th network for* Prob≈*0, for each examined molecule.*

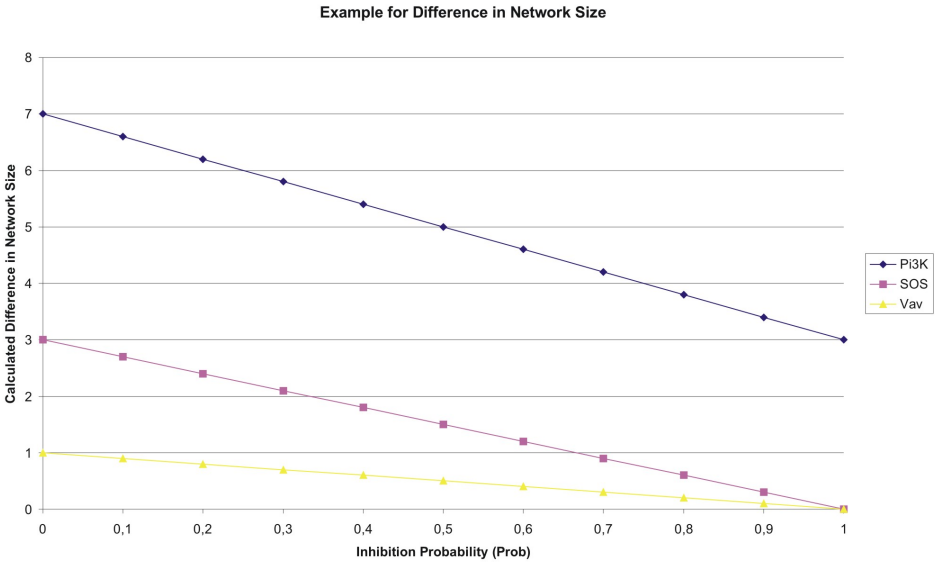
with its probability, for each molecule and each probability. As an example 'Difference in Network Size' (DiNS) for Pi3K calculates to (see values in table 2.8):

$$DiNS_{Pi3K}(Prob) = 3Prob^2 + (6+4)\left[(1-Prob)Prob\right] + 7(1-Prob)^2$$

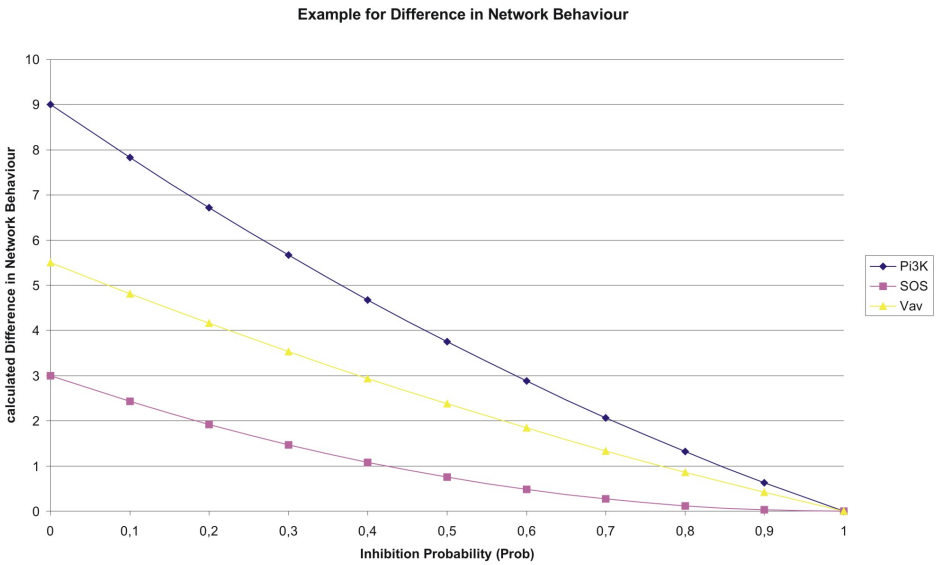
$$= -4Prob + 7$$

The results obtained this way are shown in figure 2.16. In the example all final DiNS can be calculated as a linear functions of *Prob*, in complex systems this is not necessarily the case (e.g. see figure 3.11).

Initially, when *Prob*= 0 only the network with the single DiNS= 7 exists, as *Prob* increases, this network gets less and less probable, whereas the probabilities of networks with lower single DiNS increase. If *Prob* gets closer to one the network, with missing SOS and Vav, gets more and more dominant, bringing the single DiNS closer and closer to 3. However, Pi3K has a high DiNS over the entire range of *Prob*, as can be seen by comparison with the DiNS(*Prob*) of the other molecules, shown in figure 2.16.



(a) The graphic shows the 'Difference in Network Size' in dependence on *Prob* for Pi3K, SOS, Vav, as calculated in table 2.8. Pi3K has for all values of *Prob* the highest 'Difference in Network Size'.



(b) The graphic shows the 'Difference in Network Behaviour' in dependence on *Prob* for Pi3K, SOS, Vav, as calculated in table 2.8.Again, Pi3K receives for almost all values of *Prob* the highest Difference.

Figure 2.16: Example graphs

Chapter 3

Applications

The method described in the previous chapter was applied on the well understood Lytic-Lysogeny decision circle to see if it returns valid results. Afterwards it was applied on CD34+ an K562 hematopoietic cells treated with EpoR and it identified some differences in the dynamics of these two cell lines.

3.1 Application of the algorithm on lytic-lysogeny decision in phage λ-infected *Escherichia coli* cells

The lysis-lysogeny decision is a well understood process, that among others has been analysed with a stochastic model by Arkin *et al.* [75]. The stochastic model from Arkin *et al.* has been transformed into a directed graph and the algorithm has been applied on it. The directed graph deduced from [75] is shown in figure 3.1. Arkin *et al.* state that Cro and N get transcribed immediately after infection. That is why Gene-Cro and Gene-N were set to '+' for the calculations.

Results of some examples are shown in table 3.1. The calculation shows, that there are only two stable results :

- One where CI, CI2, Gene-CI are '+' and Cro= '-'.
- One solution with CI, CI2, Gene-CI being '-' and Cro= '+'.

The state of Gene CI at the 1st time point determines which stable state will be reached. If fluctuations establish a positive feedback cycle between CI, CI_2 and Gene CI, the population continues on the lysogenic path (time point 2), else after passing a transient state (time point 2) the system stabilises at a stage with CI= '-' and Cro= '+' (time point 3), the lytic path.

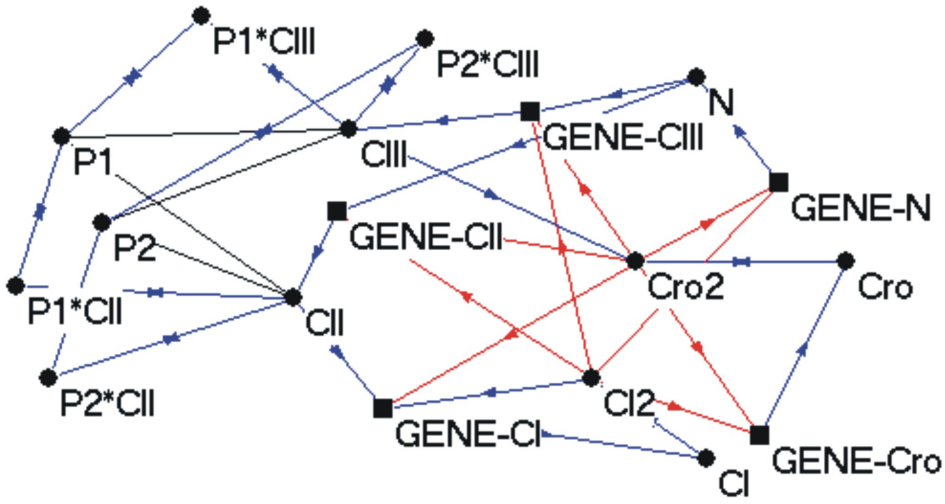


Figure 3.1: Network underlying lysis-lysogeny decision
This network was created, by transforming the ODEs found in [75], into semantic relations. Boxes, indicate genes, circles indicate molecules, blue arcs symbolise activating and red arcs inhibiting relations.

Run	CI	CI_2	Cro	Cro_2	CII	P1	P1*CII	CIII	P1*CIII	P2	P2*CII	P2*CIII	N	**Cro**	**CI**	**N**	**CII**	**CIII**
1	+	+	+	+	−	+	+	−	−	+	+	+	−	−	−	−	−	−
	−	−	−	−	−	−	−	−	−	−	−	−	+	+	−	+	+	+
	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+
	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+
	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+
2	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−
	−	−	−	−	−	−	−	−	−	−	−	−	+	+	−	+	+	+
	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+
	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+
	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+
3	+	+	+	+	−	+	+	−	−	+	+	+	−	−	−	−	−	−
	−	−	−	−	−	−	−	−	−	−	−	−	+	+	−	+	+	+
	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+
	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+
	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+
4	+	+	+	+	−	+	+	−	−	+	+	+	−	−	+	−	−	−
	+	+	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−
	+	+	+	+	−	−	−	−	−	−	−	−	−	−	+	−	−	−
	+	+	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−
	+	+	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−
5	+	+	+	+	−	+	+	−	+	+	+	+	−	−	+	−	−	−
	+	+	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−
	+	+	+	+	−	−	−	−	−	−	−	−	−	−	+	−	−	−
	+	+	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−
	+	+	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−

Table 3.1: Example results for lytic-lysogeny calculations

Because of limited space genes are indicated by bold letters. This is a table similar to table 2.5. On the long run (Time Point ≥ 3) there are only two stable solutions: One with CI = CI_2= Gene-CI= '+' and one with CI = CI_2= Gene-CI= '-'.

These results coincide well with descriptions found in textbooks:

> "Phages have a lytic life cycle in which infection of a host bacterium is followed by production of a large number of phage particles, lysis of the cell, and release of the viruses. Some phages also can exist in lysogenic form, in which the phage genome is integrated into the bacterial chromosome and is inherited in this inert, latent form like any other bacterial gene [99]."

> "The program for the lysogenic and lytic pathways are so intimately related that it is impossible to predict the fate of an individual phage genome when it enters a new host bacterium ...
>
> The same pathway is followed in both cases right up to the brink of decision. Both involve the expression of the immediate early genes and extension into the delayed early genes. The difference between them comes down to the question of whether repressor or Cro will obtain occupancy of the operators [99]."

3.2 CD34+ cells treated with Epo

The programme described in section 2 was used to to analyse how CD34+ cells react on Epo, an natural chemokine inducing differentiation of stem cells towards erythrocytes (see figure 1.1). Affymetrix time series, consisting out of four time points, were utilised as data. The 1st time point is the expression profile of untreated CD34+ cells. The other three time points are measurements made after Epo was added to the medium. The entire time course comprises several days.

The data was discredited with Micro Array Suite 5.0 from Affymetrix, keeping the default parameters. This provided the information, whether a protein was absent or present and if expression increased or decreased with time, leading basically to a table similar to table 2.4, consisting of too many genes to be shown here (interested readers will find more information in [100]).

The network derived from Transpath was modified by removing 'ABSENT' genes and proteins, having only 'ABSENT' genes encoding them (figure 2.9).

The first calculations were performed by setting EpoR= '+' and calculating results, similar to figure 2.10. Afterwards, in addition to EpoR = '+', the differences in expression of the genes were also introduced as initial values into the calculations, this is analogous to what is shown in figure 2.10. 30 simulation time points were used and the measured data was entered equidistantly, leading to a table similar to table 2.5, but consisting of approx. 900 elements, 30 Sim. Time Points and 1000 Runs.

Because the resulting networks are large (over 100 elements), just a few results will be presented. At first some groups of proteins, that have the same temporal behaviour throughout the entire calculations and thus might be something like functional modules, will be presented.

Afterwards it will be described how some proteins change their value in the network, while the cells get more and more mature.

3.2.1 Completely correlated groups of proteins

The analysis reveals 79 groups of molecules having a correlation coefficient of 1 or -1 during the entire process. Only the groups consisting of 10 or more components will be described.

AKT1-Group: In this group AKT-1, cyclin A:Cdk1, PKAc, RelA, SRY{p} and the genes IL-12p40 and ICAM-1 are correlated with each other. FOXO, Caspase-8, Ci, Ci:CBP, Cos2:Fu:SU:Ci and the gene G6PC are also correlated to each other and anti correlated to the above mentioned molecules. The components PKAc, SRY{p}, RelA, Gene IL-12p40, Gene ICAM-1, Ci Ci:CBP and Cos2:Fu:Su:Ci are connected directly with each other. AKT-1, FOXO1 and Gene G6PC are also connected. Two steps into the neighbourhood of these components were examined in an attempt to connect all of them. PDK1 that is not member of the group connects Akt-1 with PKAc. Caspase-8 is connected to Akt-1 via eNOS-NO.

The model gives the following dynamic for this group: Initially Akt-1= '+'. In the next measured time point it gets '-' to be become '+' again in the last two time points. As the molecules have either the correlation coefficient 1 or -1, the dynamic of the other molecules in this group can be deduced easily.

ER-α-Group: This group contains many genes. Grb-2:Sos, PR, R-Smads, Co-Smads and SOS are positively correlated to each other and negatively correlated to ER-α and the genes it is regulating: c-fos, Ug, NQO1, PTMA, Oxt, HSP27, SLC9A3R1 and additionally negatively correlated to NHERF-1, FTase, HePTP, MEK, AML1, RSK, RSK2 and the genes Bcl-2, βARK-1 and IL-2. Almost all of the components can be linked to each other by adding ERK to this group. The gene IL-2 has 9 upstream regulators and thus possesses many possibilities to connect to this group.

ER-α initially is '-', but EpoR turns it to '+' and it keeps this value all the time the other molecules of this group behave accordingly.

EpoR-Group: A short connection between EpoR and the other elements of this group could not be found. If there is a connection then it consists of more than two steps. All the other elements can be connected with

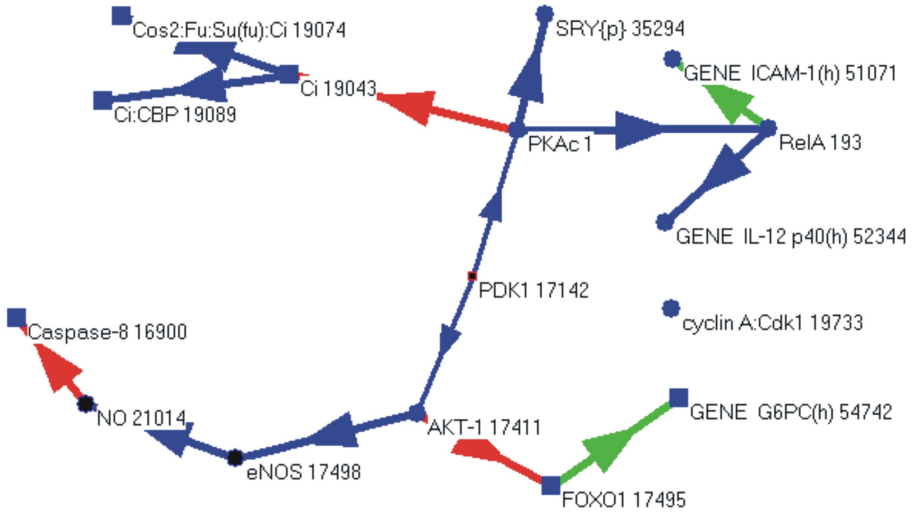


Figure 3.2: GroupAkt1

The symbolism has been changed from the one in the previous chapter. Here the colour indicates group membership, molecules which are not part of the group are black. Boxes stand for '-' and circles for '+'. Green coloured arcs are either transregulations or expressions.
Akt1 and PKAc and their downstream molecules function as a module in CD34+ cells, they are under influence of PDK1. In K562 cells Akt-1, PKAc and their downstream molecules do not have the same dynamic any more.

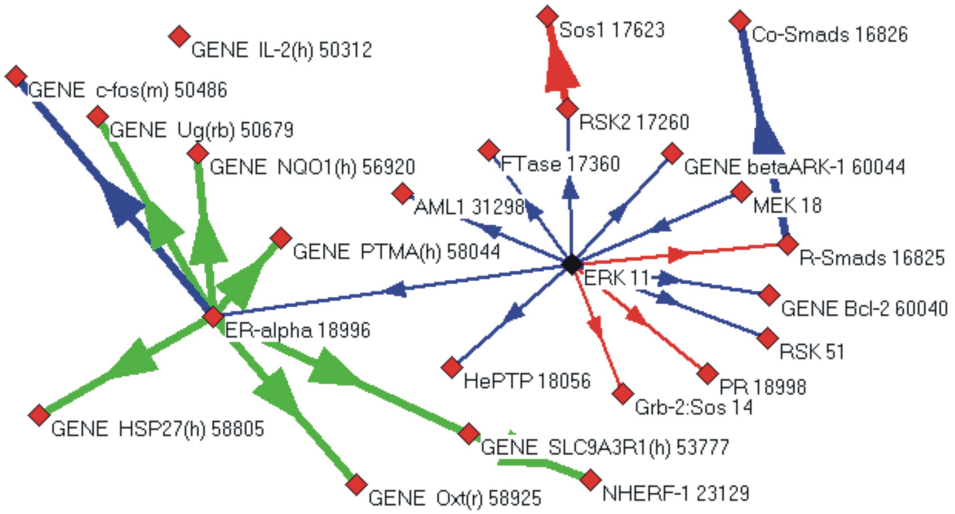


Figure 3.3: ErαGroup
In CD34+ cells ERK controls the value of Erα and the other shown molecules have the same dynamic as Erα. In K562 cells this group still exists, but the dynamic of MEK, RSK 2, SOS, AML 1, R-Smads, Co-Smads, gene Bcl-2, IL-2, βARK-1 become independent of the group dynamic.

each other by Caspase-3. CAD, CAD:ICAD, PITSLRE, Mst2, PRK-2, Mst1, PKCθ, IKK-α{p}:IKK-β{p}:IKK-γ(2) and Bmx are correlated to EpoR whereas lamin B1, CaMKIV, 14-3-3τ:PKCθ and PARP are anti correlated.

The dynamic of the EpoR group is interesting. The EpoR effect gets initially counteracted and it is after the first expression measurement when EpoR remains '+' all the time.

Ca^{2+}-Group: This group is completely connected and does not require any non-group elements for linking. The molecules at the top of the chain are Ca^{2+} and PLC. All elements of the group except RICS are positively correlated.

In most simulations Ca^{2+} is '-' at the beginning, gets '+' for the next two measurements and then '-' again for a short time to return to the value '+'. The dynamics might be truly that erratic or it might be that this group does not play any important role.

talin-Group: In this group all elements are positively correlated to each other. All elements except Cat-2 are also connected to each other and talin is at the top of the group. If the non-group element FAK1 gets added to this set, then all elements are connected with each other,

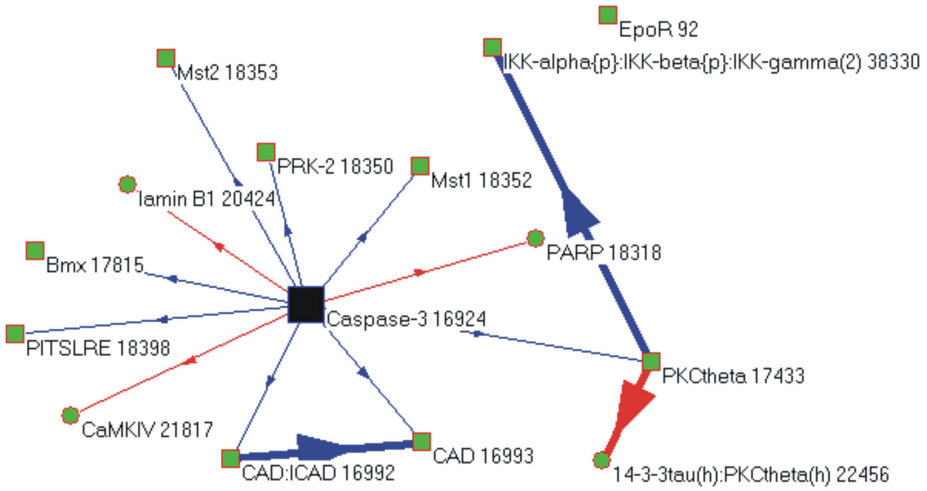


Figure 3.4: EpoR-Group
EpoR has the same dynamic in CD34+ cells like the other members of the group, although I could not find which of the many possible connections is the responsible one. In K562 cells the dynamic behaviour of this group is independent of EpoR.

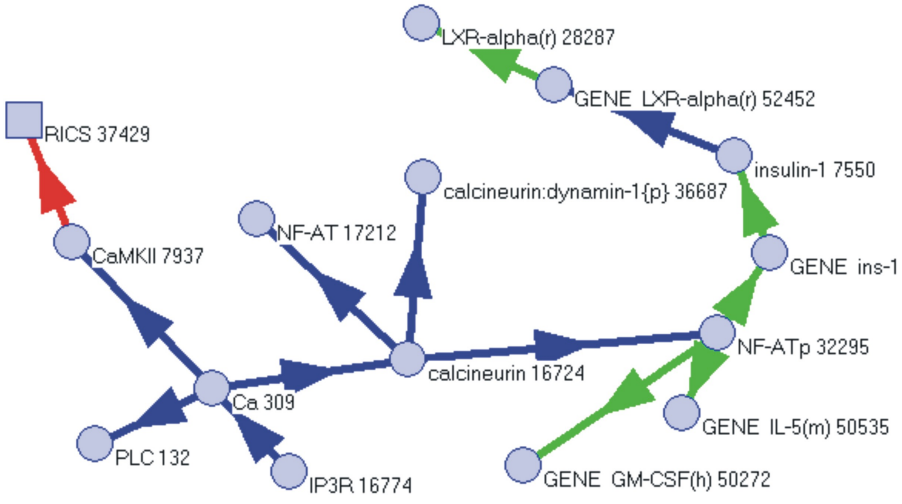


Figure 3.5: Ca^{2+}-Group
The dynamic of the proteins in this group is the same only in CD34+ cells. In K562 cells the dynamics are decoupled.

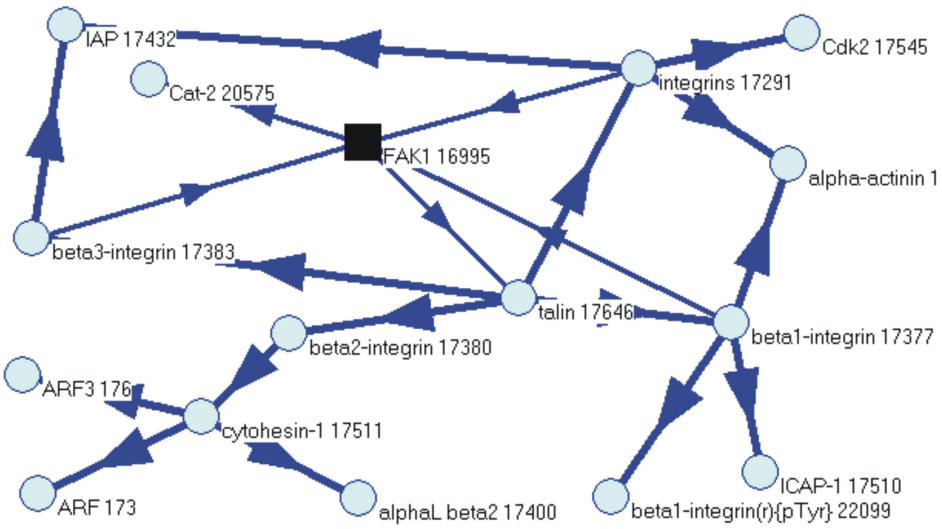


Figure 3.6: talin-Group
This group is found almost unchanged in both cell types.

but as FAK1 influences talin and gets itself influenced from inside the network by β1-integrin and β3-integrin there is no real top molecule in this group.

Talin in the simulation is '+' at the beginning, gets '-' between the 1st and the 2nd measurement point and then gets '+' again.

IL-6-Group: This is a big group consisting of 41 elements, approximately half of them being genes. At first sight this group appears to consist of unconnected elements, but most of the group members can be connected easily by adding the non-group element IL10, then, only Gene HSD17B1 is not connected to the network. Adding Sp1, that is a regulator of the IL10 Gene and Gene HSD17B1, results into a connected network. The group consists of correlated and anticorrelated elements. The image gives an overview of the topology and the group members.

IL is '-' at the beginning. With the 3rd measurement point it gets '+' and remains '+' throughout the differentiation process.

c-Kit-group: It was not possible to connect all elements of this group to each other within two steps. Half of the elements in this group are depending on c-Kit and show a +1 correlation to it.

C-Kit is initially '-', but after the 1st expression experiment remains '+' all the time.

IFN-γ-group: Only few elements of the group are connected with each

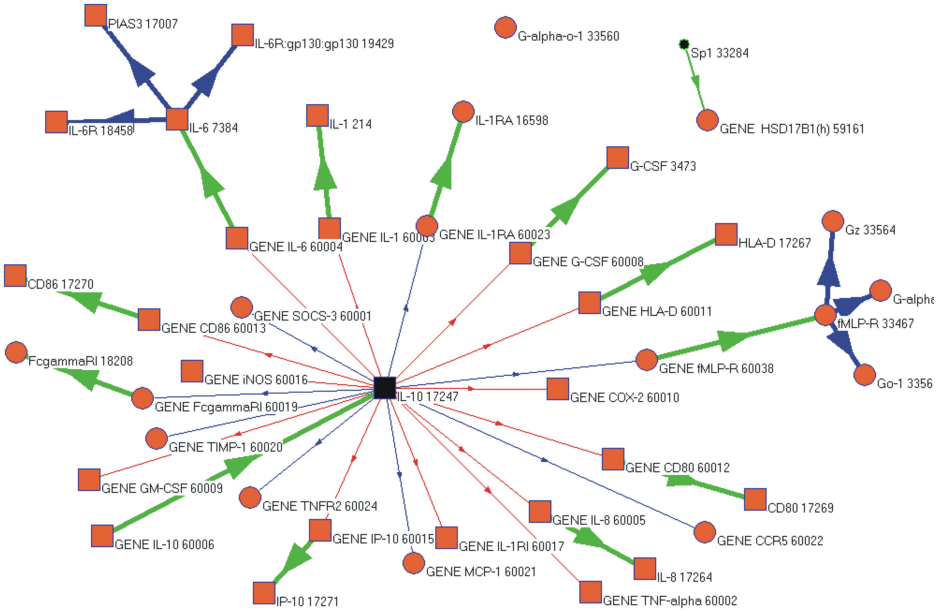


Figure 3.7: IL-6-Group

In CD34+ cells IL-10 controls the dynamic of shown genes and proteins. In K562 cells this group does not exist, leading to the conclusion that dynamic of other proteins than IL-10 determine the dynamic of parts of this group.

other.

IFN-γ gets '+' by influence of EpoR and remains activated all the time. The other parts of this group behave according to their correlation coefficient.

IFN-β-group: The elements of this group cannot be connected within two steps.

The elements of this group having a positive correlation to IFN-β are initially '-' or not expressed. After the last measurement they receive the value '+'.

3.2.2 Investigating the network for distinct timepoints

The dynamic during the 1st four time points, simulating the events taking place from the initial state until the 1st measured point can be grouped into distinct groups each having its distinct dynamics.

Big group with 373 elements: The previously described groups *IL-6-Group, c-Kit-group, Ca^{2+}, talin-group, IFN-β-group* are part of this subgroup, hence the temporal behaviour of the positive group members is: '+', '+', '+', '+', in other words the members of this group achieve a value very fast and do not change it.

- The model calculates, that PU.1 is '+', leading also to a '+' for the genes Sfpi1, ISG15, Acp5 and IL-12 p40. Pure IL-12 is '-' due to IL-10 being '-', which is an activator for IL-12 gene expression. The value '-' of IL-10 can be ascribed to the '-' value of its gene. PU.1 might be '+' due to the '+' of Sfpi1, leading to PU.1 = '+'.
- AP-1 = '-', this leads to a '+' for IL-2, IL-3, IL-5, c-fos, Rbp, p53, c-myc, SPRR1B, ins-1, GM-CSF and a '-' for IL-12p40. The molecule responsible for AP-1 = '-' could not be identified. Possible candidates are c-Jun and Ref-1. ILK can be excluded as explanation, because it is '+' and is linked with an activating reaction to AP-1.
- NF-ATp= '+' leads to expression of some previously mentioned genes, like ILs and GM-CSF. Additionally, expression of ICAM-1 is fostered by NF-ATp. A possible signalling chain explaining the value of NF-ATp is CD19 - BCR - FCγRIIB - SHIP - PIP2 - IP3 - IP3R - Ca^{2+} - calcineurin - NF-ATp. All elements in this chain are '+' and are linked via activating reactions.

EpoR containing group with 81 elements: This group contains the previously mentioned group *IFN-γ-group* and *EpoR-group*. Thus the dynamic of positive members is given as '-', '+', '+', '+'. In this group the model identifies among other things the following relations:

- EpoR gets '-' because SOCS-3= '+'.
- PRK2= '-' leads to PDK1= '-'.
- JNK3= '-' leads to NF-ATc= '+'. NF-ATc sets the values of multiple genes like Egr3, IL-4 and IFN-γ, all of them being members of the *IFN-γ-group*, to '+'. Proteins produced by the IFN-γ-gene lead to a '+' for CRkII via IFNRII= '+'. In addition to this IFN-γ might set STAT3 to '+', that subsequently leads to CSN2= '-'.

Differences between the 1st and 2nd timepoint

Because the network is big the 2nd network will be compared to the 1st one.

- The comparison shows, that EpoR now is '+'. The explanation for this is the '+' of c-Kit, being an upstream molecule of EpoR. The model cannot calculate the reason for c-Kit= '+', possible candidates are PKC, SHP-1 or SHP-2. The model is not able to identify clearly downstream effects of EpoR. This is likely to be explained by the many possible downstream reactions of EpoR.
- Possibly, EpoR among other things supports NFATp= '+' via Epor - PLCγ - IP3 - IP3R - Ca^{2+} - calcineurin - NF-ATp, this might be necessary, because the coupling between CD19 - BCR - FCγRIIB - SHIP - PIP2 is weak during the 2nd time point, this means the correlation between the values of these molecules is low.
- Talin receives the value '-' mediated via '-' of Akt-1 -¿ eNOS -¿ NO - Caspase-3 - FAK1 - talin leads via integrins to '-' of ERK, Abl, CDk2 and IAP. Erk via p70S6K sets PkFB-2= '-', which is also influenced by Akt-1 and RSK2.
- PKAc= '-' leads among other things to a '-' for HNF-6 and subsequently to a '-' for HNF4A and G6pc, additionally, PKAc transforms via RelA ICAM-1, p53, IL12 p40 and MEFH to '-'. The model suggests that PDK1= '-' might be a reason for PKAc= '-'. The value of PDK1 might be influenced by Abl= '-' and PRK2= '-'.

Differences between the 2nd and 3rd timepoint

- The difference network shows a positive feedback- cycle consisting of Abl - PDK1 - AKT-1 - eNOS - NO - Caspase-3 - FAK1 - talin - integrins - Abl, except Caspase-3 all members of the cycle are '+'. Caspase-3 gets '-' by NO, thus Caspase-3 inhibitory strength on FAK1 gets reduced, because of this the cycle is a positive feedback cycle. The cycle gets further stabilised by PRK2 - PDK1. This cycle leads

to Transcription Factor RelA= '+', via PKAc= '+', resulting in '+' of p53, MEFV, IL-12 p40, ICAM-1.

- Epo= '+'.
- STAT3= '-'. The inhibition is done by IL-6, experiencing a '+' expression during the observed time. STAT3= '-' is further supported by PKR.
- AP-1= '+', resulting in IL-4 and GM-CSF = '+'. The model is not able to identify the source of Ap-1= '+'. Possible sources are IL-12 p40 and Ref-1.

Differences between the 3rd and 4th timepoint

- Jak2= '+' and Jak1= '-' results among other things in STAT6= '+' and STAT5= '+'. The model is not able to give a clear answer on why Jak1 is '+', possible candidates being able to perform this task are: IL-10R, IFNRI, IL-2R complexes and other receptors. The calculations suggest a role of CTLA-4, InsR and SOCS-3= '+' in transforming Jak2 to '+'.
- The calculations identify a network starting from PIP2= '-', that connects via IP3 to the absolutely correlated network, called Ca^{2+}-Group initially. This group interacts with elements of the IFN-β-group via NF-ATp and Rac1, leading to the speculation that the IFN-β-group might be a subgroup of the Ca^{2+}-Group..

3.3 K562 cells treated with Epo

3.3.1 Completely correlated groups of molecules

Analogously to the treatment of CD34+ cells , absolutely correlated subsets of elements in K562 cells were found and afterwards compared to the results of CD34+ cells. The AKT1, Ca^{2+}-Group, IL-6-Group, c-Kit-Group, IFN-γ-Group, IFN-β-Group cannot be found in K562 cells.

ER-α-Group: In K562 cells this group is smaller than for CD34+ cells. The elements MEK, RSK2, SOS, AML1, R-Smads, Co-Smads and the genes Bcl-2, IL-2 , βARK-1 are missing.

EpoR-Group: EpoR itself is not element of this group any more but gelsolin and α-fodrin get added to this group.

talin-Group: This group is almost similar to that one found in CD34+ cells. Still one molecule FAK1 cannot be found in this group.

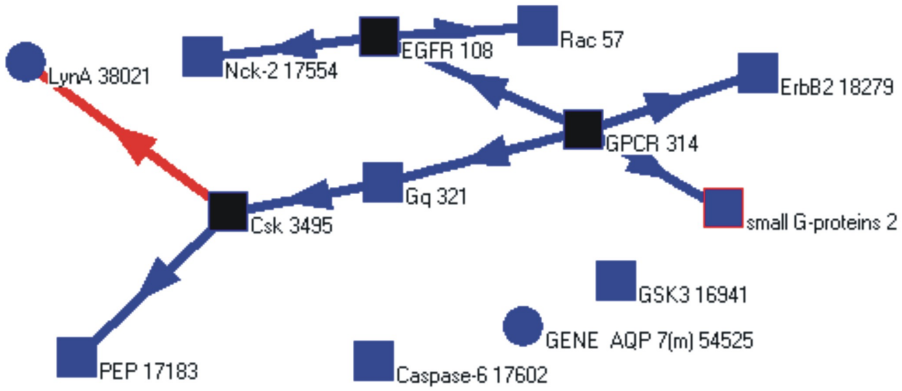


Figure 3.8: GPCR-Group
This group is found only in K562 cells.

GPCR-Group: This group is found only in K562 cells. It consists of the elements PEP, GSK3, ERbB2, small G-proteins, Rac, Nck-2, Caspase-6, Gq, LynA and Gene AQP 7. If Csk, GPCR and EGFR are added to the group, then many elements are connected with each other.

3.3.2 Investigating the network for distinct time-points

Abl-containing groups with 388 elements: The previously mentioned ER-α-Group, the talin group and the GPCR-group are part of this group. All elements of this group achieve their final value fast and keep it throughout the 1st four time points.

- Abl= '+' and effects many elements of this group.
- The value of Abl is supported by pRb= '-', an inhibitor of Abl. The reason for reduced pRb activity cannot be calculated exactly by the model. One possible cause might be RB gene= '-', mediated via ATF-2.
- ERK being '-' might be important to set genes regulated by ER-α to '-'.
- Interestingly EpoR has increased activity in these calculations but the behaviour of most of its downstream elements is dominated by another process. Therefore, the elements that get activated by EpoR, for example Jak 2, Stat 5, Vav, Grb-2, Shc, Pi3K and PLCγ, have a reduced activity. The reason might be reduced activity of c-Kit, due to inhibition by SHP-1 or SHP-2.
- EpoR leads to Fes= '+' and subsequently to STAT5a= '+'.

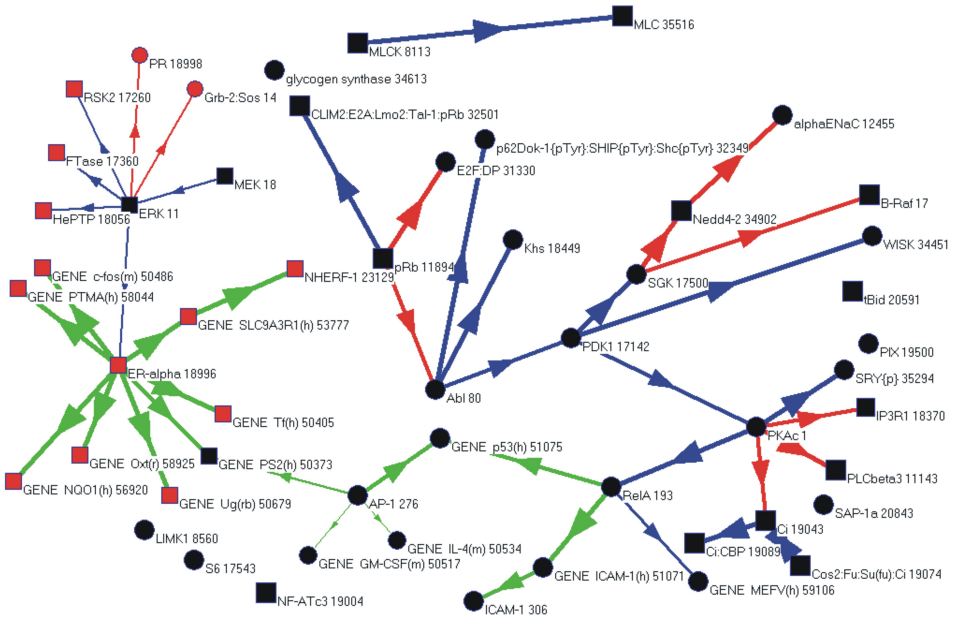


Figure 3.9: Network around Abl

Differences between the 1st and 2nd timepoint

- ATF-2= '+' leads to Gene RB= '+' and subsequently to Abl= '-'. PDK1 and AKT-1 propagate the value '-' further into the network. This results among other things in Caspases-3= '+' and Caspase-6= '+'.

- The network has a positive feedback cycle solely consisting of activating reactions that are build of Erk, FTase, Ras, Raf , MEK, Erk. The cycle leads via multiple steps to a '+' of the Genes Bcl-2, TNF-α, SLC9A3R1, Ps2, c-fos, Tf, TGF*alpha*, PTMA and HSP27.

Differences between the 2nd and 3rd timepoint

- Gene RB= '-' reversing all effects described in the first part of the previous section.

- The previously discovered cycle is found again in this network, but with inverted sign. The coupling between cycle members is much weaker, therefore the cycle might not be active any more.

3.4 Sensitivity analysis

The method described in section 2.3.2 was applied on the measurements for untreated CD34+ cells to identify the molecules, that have a high influence on overall system dynamics. The Transpath network is adjusted to the CD34+ network by removing unexpressed proteins.

As exemplified in figure 3.10 for PDK1 the effect on the difference in network size induced by expression/not expression of PDK1 depends strongly on the inhibition probability. If the inhibition probability is extremely low alternative connections lead to a narrow distribution of possible network sizes and most networks are big. For intermediate inhibition levels the distribution is very broad and gets narrow again for high inhibition probabilities. This is consistent with the example shown in section 2.3.2, where for $Prob \approx 1$ or $Prob \approx 0$, the probabilities increase strongly in favour of one of the extreme network, whereas intermediate values of *Prob* increase the number of different networks, that have to be considered.

The average 'Differences in Network Size' and average 'Differences in Network Behaviour' for the treated elements for different inhibition probabilities are shown in figure 3.11. Most curves are monotonous, but this is not the case for all of them. The results can be explained by two effects:

Redundancy : For low inhibition probabilities many molecules are connected with each other among multiple, different paths, if a single molecule (e.g. the examined one) is removed, then the network, does not change drastically, because of alternative connections. This leads to low Values of 'Difference in Network Size' and 'Difference in Network behaviour'.

Reachability : A molecule far away from EpoR, requires a signal that has passed other molecules. If this molecules are completely inhibited, then the signal does not reach the molecule far away from EpoR, hence molecules afar from EpoR, require low inhibition probabilities to obtain high values.

The relation between 'redundancy' and 'reachability' determines, where the maximum of a molecule is found in figure 3.11.

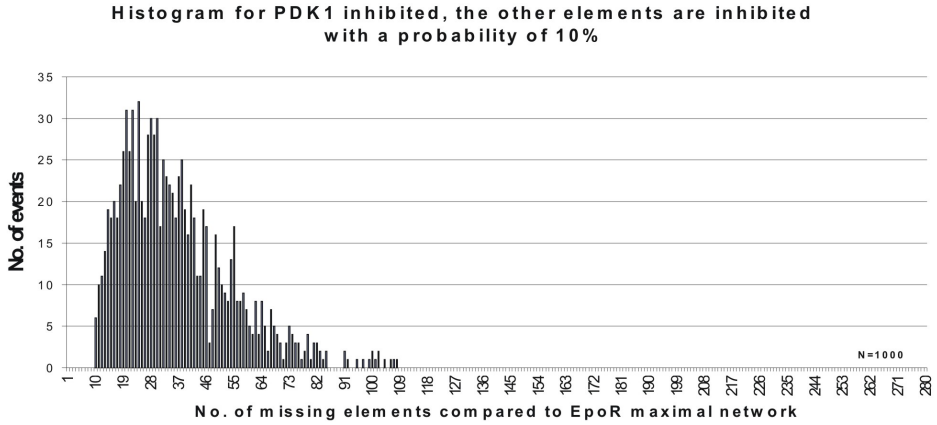
Some elements like Pi3K, PIP3, Akt receive high values. This are the molecules that are considered to have strong influence on EpoR induced hematopoietic differentiation. Differences between the network obtained for Cd34+ cells and K562 cells are small (figure 3.12). This may indicate, that changed dynamic is more important for CML, than changed topology.

Ranking the proteins by the method described in 2.3.2 shows that Pi3K, Lyn and PIP3 have the highest values as shown in table 3.2. As described in section 2.3.2 the Factor used to rank the molecules, is basically a number showing 'Difference in Network Size' and 'Difference in Network Behaviour'

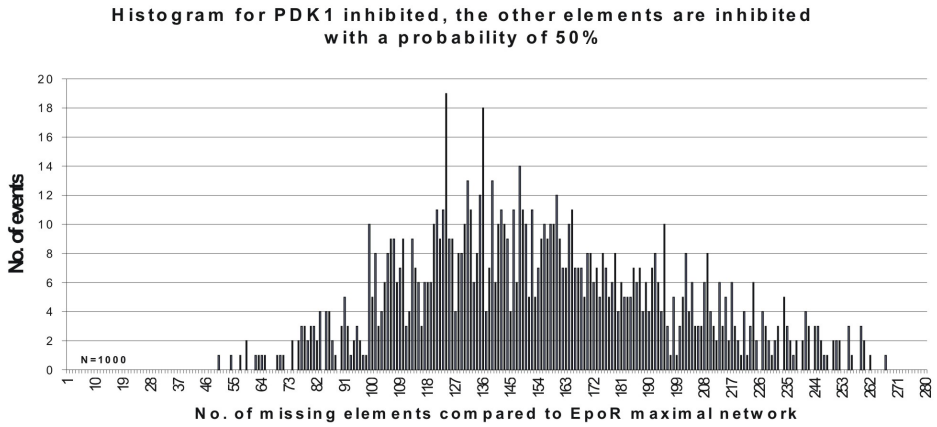
Molecule Name	Factor for NW Size	Factor for NW Behaviour
AKT	6.07	2.92
PIP3	4.22	3.79
Pi3K	4.17	3.94
Lyn	2.91	4.64
Caspase-3	2.24	2.26
Gβ:Gγ	2.07	1.25
ERK	2.03	2.32
Syk	1.96	2.65
DAG	1.80	1.50
Jak1	1.54	1.67
Fyn	1.26	1.62
Cdc42	1.22	1.63
Ras	1.16	2.03
p38	1.13	2.13
RhoA	1.12	1.43
SLP-65	1.09	1.30
Src	1.04	1.18
SLP-76	1.01	1.17
p38α	-	1.04

Table 3.2: Ranking of molecules
The values obtained for 'Difference in Network Size' and 'Difference in Network Behaviour', for each value of Prob*, are summarised into single value, shown here. Most molecules with high 'Factor for NW Size' have also a high 'Factor for NW Behaviour'.*

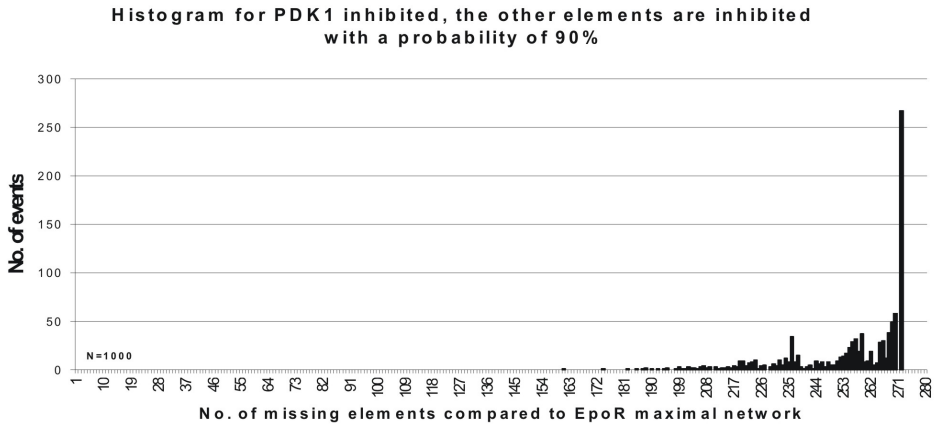
averaged over all probabilities *Prob*, divided by the average standard deviation, for each molecule.



(a) Histogram

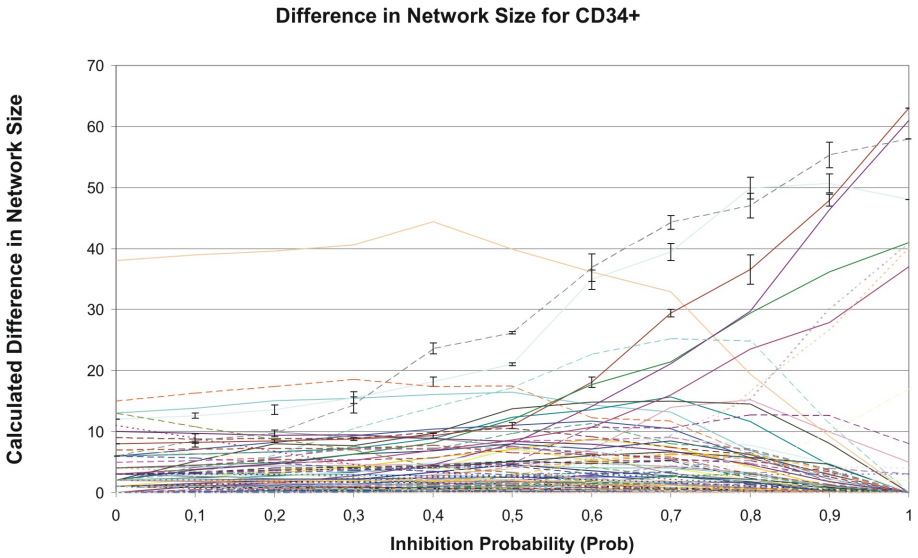


(b) Histo

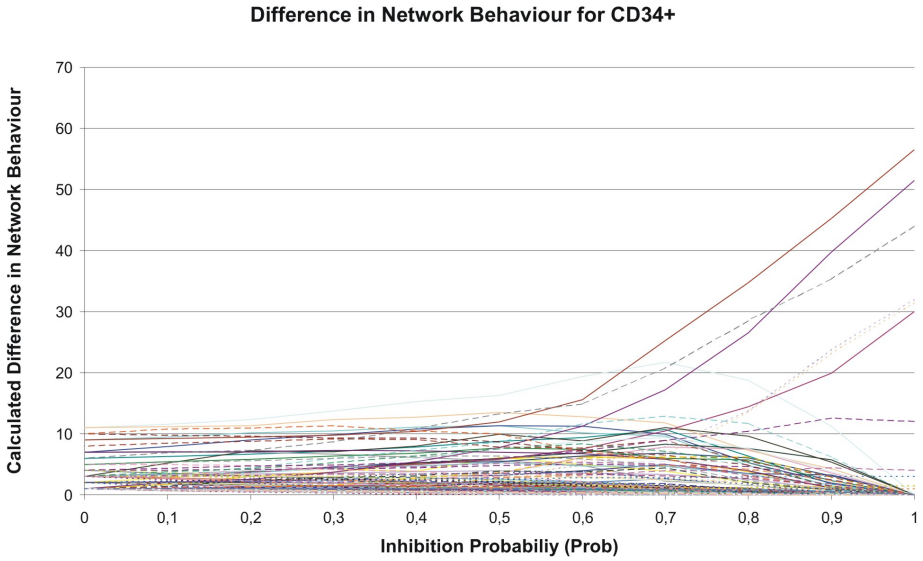


(c) Histogram

Figure 3.10: Histograms for PDK1
The histograms show, that the network distribution, approximated by the distribution of network sizes, depends strongly on inhibition probability
Prob.



(a) The figure shows that most calculated 'Differences in Network Size' accumulate at the lower part of the graph. Few outliers can be found easily: Akt has at low probability of inhibition a value of 39 then this value increases till an inhibition probability of 0.4 is reached and finally declines for high inhibition probabilities. In contrast to this Lyn, Syk, PIP3 and Pi3K receive high values for high inhibition probabilities and lower ones for low ones.



(b) Two groups can be distinguished easily, one group consisting of Lyn, Syk, PIP3 has high values for high inhibition probabilities that decrease fast. The other group consists of Pi3K and DAG, whose values peak at an intermediate inhibition probability.

Figure 3.11: Differences for CD34+ cells.

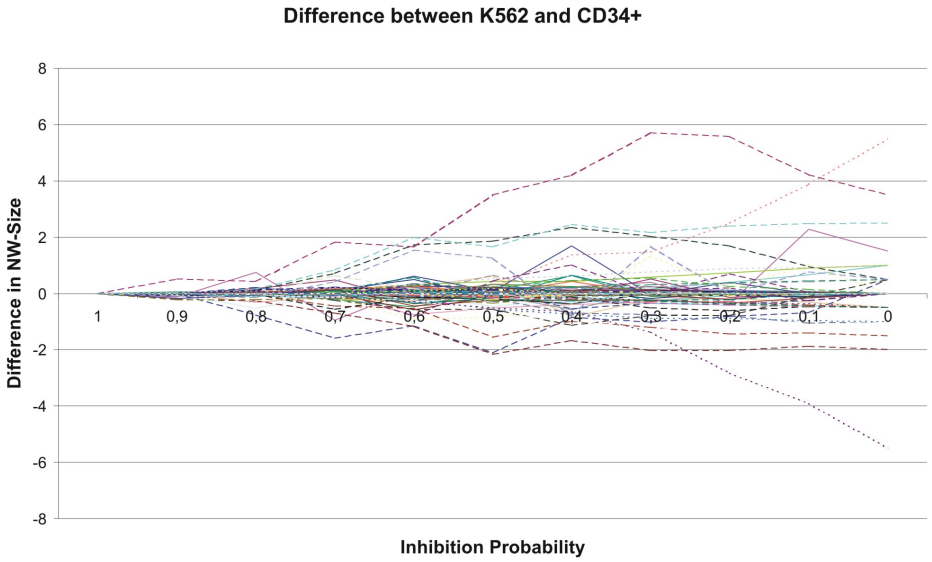


Figure 3.12: Difference between K562 cells and CD34+ cells
The calculated differences in network size for treated elements between the Cd34+ and K562 networks are rather small. Only for c-Kit, cyclinB:Cdk1, PDK1, FAK1, Shc, Bax and Bad(pSer):14-3-3τ, the absolute value exceeds 2.

Chapter 4

Discussion

The aim of this work was to find a method, that is able to work with a few measured time points and that takes cyclicity and positive feedbacks into account.

The approach described in chapter 2 is able to do this.

Few time points: The problem arising from the low number of measured time points is mainly solved by allowing only the reactions that are found in the database. This restricts the solution space strongly. Without this restriction the 33000 measured genes, assuming that every gene can theoretically interact with any other gene and itself, would have 1089000000 theoretical reaction possibilities and all networks that could be build out of these reactions by removing some of them would have to be checked against the measured data to find the best one. It is obvious that a time series consisting of 4 or even 10 measurements does not contain enough information to find the right network. If the usual restrictions, like limiting the number of interaction partners to for example three, were used, then the same problem would still remain, because the number of combinations for networks would still be huge.

In Transpath there are 23062 reactions, thus 47220 times fewer reactions than theoretically possible. Because the number of possible networks increases exponentially with the number of possible reactions [95] the number of required measurements decreases drastically.

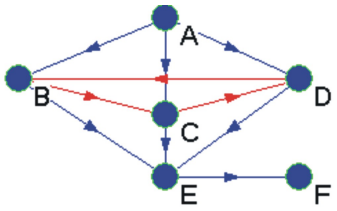
The advantage is that not so many measurements are required. The drawback is, that one has to rely on database information and one might get wrong results if reactions in the database are wrong. Because the algorithm tries to find the most relevant reactions and thus ignores some reactions false positive reactions to some degree are not that critical. More critical are false negative reactions, because the programme does not search for them.

Cyclicity: Almost all current Bayesian approaches that infer networks from data ignore cycles even in differentiation networks despite the knowledge, that positive feedback loops are important for these networks [73, 26, 94, 93, 61]. Feedbacks are neglected because there are no theoretical means to deal with them in the Bayesian framework. As other approaches try to infer structure from data they have a big search space $r(n) = \sum_{i=1}^{n}(-1)^{i+1}\binom{n}{i}2^{i(n-i)}r(n-i)$ [101], that cannot be exhaustively searched, therefore, they need a decomposable system that can be treated with the Bayes approach and thus cannot consider feedback loops.

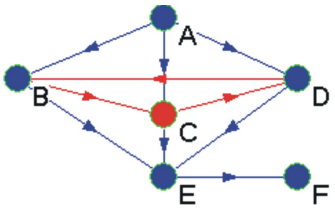
Because, in the method used in this work, the search space is much more confined, it is possible to calculate many networks and it is theoretically feasible to find the best solution by scoring the resulting network against the data. Because practically just 1000 calculations are used, one cannot be really sure, that the best solution is found by the calculations. This problem is solved by taking approx. the 20 best solutions and comparing their similarities. Then the reliability of feature can be calculated, by the number of its occurrence. For example, if A activates B in all solutions, then this result is reliable, if B activates C in 5 of 20 solutions, then this result is less reliable. The approach allows to identify the well known facts and separate them from not well known parts of the network and is able to deal with cycles, because decomposition is not necessary. It has the disadvantage that molecules belonging together might get separated due to many degrees of freedom and therefore uncertainty in the system (see figure 4.1).

Reactions exceeding thresholds: Usually, Bayesian networks are learned by looking for correlations between elements among different experimental conditions or time points. This approach assumes, that relations between elements remain constant. The shown approach, by contrast allows, changes in network structure, because it does not use the experimental data to calculate a priori probabilities, but instead calculates a network for one time point, that then is allowed to change using the relations in the database, thus producing a network for another timepoint that than is compared against the calculated data. By connecting the networks this way changes between the dependencies of variables over time can be considered.

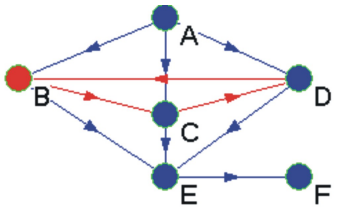
The approach described in chapter 2 to calculate the network uses the discrete values ‘0’, ‘+’, ‘+’, ‘uk’,‘PRESENT’ and ‘ABSENT’. The values ‘PRESENT’ and ‘ABSENT’ are used to determine network topology and are not further checked by the algorithm. The values ‘0’, ‘+’, ‘+’, ‘uk’ are used to determine how the network works as a whole system.



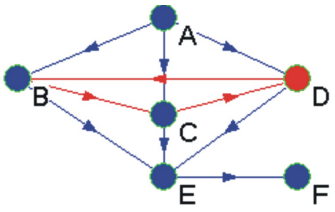
(a) If this directed graph and the information A= '+' and F= '+' are given, then the algorithm could identify and the following solutions as good: The one shown here where all molecule are '+', ...



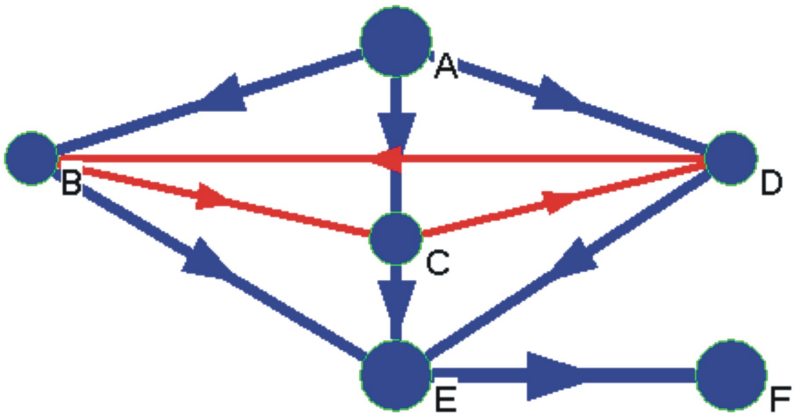
(b) ... this solution where the interaction C – D is assumed to be dominant, ...



(c) ... the solution where D – B and ...



(d) ... this solution.



(e) All four solutions have the same score and therefore are equally well. The algorithm knowns for sure, that A = E = F = '+', the values for B, C, D are in 3/4 of the cases '+' and 1/4 of the cases '-', this uncertainty is visualised by the smaller arcs sizes and the smaller node sizes. If the algorithm shows only absolutely certain values and interactions then B, C and D are removed from the graph, leading to an unconnected sub-graph.

Figure 4.1: Uncertainties in connections and values

A false 'PRESENT' is not that critical because it is just increasing search space and reduces the amount of definite conclusions. False 'ABSENT' measurements are a source of error. In this work the focus was on developing a method and therefore, this error source was omitted for simplification reasons. The resulting mistake could be small, because it was shown that in technical replicates made with Affymetrix chips the consistence of results is higher than 95% [102].

However, the problem of false 'ABSENT' calls could be solved in two ways.

1. The first and easiest solution is to leave all proteins in the network independent whether they are expressed or not, this would result in an increased number of possible solutions of the network and thus require more measurements.

2. The second solution is to use different topologies of the treated network, based on the 'ABSENT'-call information and for example, also use a number for the probability of this constellation. Then this probability and the score that was already used to find the best networks would require a binary operation (e.g. multiplication, addition) to determine the complete score.

The discrete approach is justified, because it is believed that the hematopoietic differentiation is a threshold based process that, after a decision was made, becomes self-perpetuating. Therefore, it is sufficient to identify the subsystem being responsible for this self-perpetuating effect. Because only three real values '+', '-', '0' and a helping value 'uk' are used, complex multistep processes cannot be calculated by the algorithm. However, the principle used can be expanded to a system containing more discrete values. The drawbacks of this would be that the search space increases and, therefore, the measurements required to restrict the system increase and the precision of the measurements of the protein status must also increase. It does not make much sense to use many discrete values if the measurements can give only the information if a value is 'IncExpr' or 'DecrExpr'.

Discrete systems have the advantage that they require less information than continuous systems and that their results can be interpreted better. But they have also drawbacks: For example, the model cannot calculate what happened before a self-perpetuating cycle was established, the simulation basically starts after the decision on the cell fate was made.

The obtained results are big networks, because the programme leaves all reactions with high correlation coefficient and a consistent value in the network. A possible simplification would be to change the algorithm in such a way that only the minimal network is shown. The disadvantage would be that potentially important molecules could be omitted. If enough data

is available the number of possible solutions decreases and the algorithm would remove inconsistent/unimportant molecules and reactions.

Extraction of functional subnetworks: The identification of functional subnetworks, described in section 2.2 tries to identify the self sustaining structures in biological networks. This makes sense, because it is believed that stochastic effects, that get stabilised, are important for differentiation.

The model works well in the lytic-lysogeny example, where statistic effects are known to lead to a decision. The calculations identify two stable solutions, the lytic and the lysogenic, proving that the approach is able to identify important behaviours of a system based solely on topology and some starting conditions.

The same approach is used on erythroid differentiation, however, the solution space is much bigger here, therefore, the calculated networks are filtered against the experimentally obtained data, thus only self-perpetuating structures, that match biological data well, are considered. The obtained networks are still too big, therefore, they are subdivided in smaller networks consisting of elements, which have the same dynamic in the calculations. The results can be tested for biological validity by making measurements that check if all molecules belonging to one group have the same dynamic of activity, as predicted by computations. If this is the case then the method could be used to identify molecules having the similar activity dynamic in unknown systems. Additionally, molecules with the same dynamic could be members of one feedback-cycle, having importance in differentiation processes or more generally being part of one and the same functional module.

From this point of view it is interesting to note, that a number of groups found in CD34+ do not exist in K562 cells. Because elements having the same dynamics over the entire time period of differentiation are put in one group, absence of a group means, that these elements do not have the same dynamic any more, although they might still be part of the entire network. The missing groups are: AKT1-Group, Ca^{2+}-Group, IL-6-Group, c-Kit-Group, IFN-γ-Group, IFN-β-Group. Assuming that these identified groups of proteins, having the same dynamics throughout all calculated time points, are essentially something like 'modules' or 'functional units', then absence of these groups would suggest non-functional modules. Under this hypothesis, absence of the IFN-β-Group in K562 cells would indicate, that this group might be malfunctioning in these cells. Interestingly, IFN-α is used as a therapy for CML patients, presumably by stimulating immune response[103, 104]. It may be important that the leukaemia cells themselves are not stimulated. IFN-α and IFN-β have similar functions, their sequence difference "may possibly cause different responses to various inducers, or result in the recognition of different target cell types" [105, 106]. It may be important that under this situation activity of leukaemia cells is not

stimulated, which would be coherent with a malfunctioning of this group.

Malfunction of other identified groups might be possible, if their presently known functions are considered:

- IFN-γ whose interaction with other molecules is perturbed in K562 cells is a protein with important immunoregulatory functions, it can increase antitumor effects of type1 interferons [105].
- The model calculates Akt deregulation. It is known that Bcr-Abl, a fusion-protein found in many leukaemias (see section 1.1.3) activates the Pi3K pathway [107] resulting in Akt deregulation [108] as calculated by the model. Activation of Akt enhances mitosis [109].

The other mentioned malfunctioning modules identified by the calculations and the additional GPCR-group found only in K562 cells, might also play important, but presently unknown roles in chronic myeloid leukaemia.

Sensitivity analysis: Jeong *et al.* [110] emphasize the importance of hubs for maintenance of network connectivity and functionality. Jeong *et al.* assume a known and stable topology. In the network, which is developed from Transpath data, the used network consists of reactions, that were observed in some systems under certain conditions. In some cases the conditions and the systems are different from those used in the experiments made for this work, thus some reactions will not be possible in reality, because, for example, the parameter-range of the experimental system is such, that proteins are inhibited or some theoretically possible reactions must not be performed due to thermodynamic constraints. These uncertainties are considered by building multiple alternative and possible networks and determining the molecules that are important for the entire set of constructed networks. The different networks are built by assuming an inhibition probability for proteins that have the potential to be inhibited. This approach emphasizes the importance of elements that are close to elements, that are known to be changed in the network. Elements far away from nodes, whose status is not known receive lower importance, despite their possibly high connectivity numbers.

This can be seen in figure 3.11 for Akt, that needs a lower inhibition probability to get connected to EpoR.

The calculated numbers consider a variety of factors:

1. The proximity to EpoR: Molecules far away from EpoR are connected via molecules that are completely inhibited with a certain probability and therefore do not get reached in every simulation. As they are not part in every randomly generated network their overall influence on the average network behaviour is reduced.
2. The number of interconnections: Highly interconnected molecules have a higher probability to achieve high values. But a high number of

interactions is not sufficient to achieve a high difference, because its effects can be counteracted, by alternative paths and the proximity to EpoR.

3. The value labelled as Difference in Network Behaviour takes that kind of information into account and thus gives information beyond the topology.

Current biological knowledge is consistent with results obtained by the calculations. As shown in table 3.2, EpoR induced hematopoiesis in CD34 cells should be sensitive to changes in AKT, PIP3, Pi3K, Lyn, Caspase-3, Gβ:Gγ, ERK, Syk, DAG, Jak1, Fyn, Cdc42, Ras, p38, RhoA, SLP-65, Src, SLP-76 and p38α, where sensitivity for AKT, PIP3, Pi3K ... is much higher (high ranking in table) than for SLP-76 and p38α (low ranking in table).

As published Akt, Pi-3K ERK, p38α and Ras are deregulated in CML cells [111, 112, 42, 113, 114, 43, 115]. As far as I know the other identified molecules are not associated mainly with hematopoiesis. Many of them are well known cellular regulators, also at present not connected with hematopoiesis, some of them could play a role in hematopoiesis, due to their presently known cellular functions:

Src, Lyn, Fyn : "The members of the Src kinase family are expressed in a wide variety of tissues, but some of them such as ... Lyn are found primarily in hematopoietic cells ... Our results indicate that Fyn and Lyn are efficiently cleaved in their unique region in hematopoietic cells undergoing apoptosis ... Thus, cleavage of Fyn and Lyn during induction of apoptosis represents a new mechanism for the regulation of Src kinases that may have important functional and physiological consequences"[116].

Caspase-3 : "Caspase-3 is a member of the cysteine protease family, which plays a crucial role in apoptotic pathways and can be activated by diverse death-inducing signals, including the chemotherapeutic agents ..."[116].

Syk : "Syk is a protein tyrosine kinase, that is widely expressed in hematopoietic cells. It is involved in coupling activated immunreceptors to downstream signalling events that mediate diverse cellular responses including proliferation, differentiation and phagocytosis" [116].

DAG : "Dystroglycan is a laminin binding component of the dystrophin-glycoprotein complex ... Forms part of the dystrophin-associated protein complec (DAPC) which may link the cytoskeleton to the extracellular matrix ..." [116].

Jak1 : "Janus kinase 1 is a protein tyrosine kinase ... Numerous aspects of lymphoid and myeloid cell functions are controlled by a group of

ligands termed cytokines, all of each signal through a related set of receptors. All such receptors are associated with one or more members of the JAK family." [117].

Cdc42 : "The protein encoded by this gene is a small GTPase of the Rho-subfamily ... Cdc42 regulates anchorage-independent growth and is necessary for Ras transformation" [116].

RhoA : "Belongs to the small GTPase superfamily ... Regulates a signal transduction pathway linking plasma membrane receptors to the assembly of focal adhesions and actin stress fibres" [116].

SLP-65 : SLP-65 is transmembrane receptor protein tyrosine kinase adaptor [118] and plays a role in hemocyte development [116].

Considering the size and complexity of the network downstream of EpoR (see Figure 1.4) a mathematical approach to obtain an overview is necessary. The ranking of molecules obtained by the symbolic simulations is an important aid to experimentalists, showing them the possibly important molecules of a part of an network in a cell and in this way giving them an overview and orientation. As the molecules identified as relevant by the model are likely to contain a lot of information on overall system status, they are a good starting point for measurements and to perform quantitative simulations of the over-all system. The importance of the calculated molecules is arising because of two reasons:

- The molecules play, as shown by comparison to known knowledge, an important role in hematopoiesis.
- Some identified molecules are hubs, with many connections in the network and thus very likely, contain integrated signals of overall network status.

The function of some molecules identified by the model, like Lyn, Syk, DAG ..., in hematopoiesis is at present not known. These molecules might be important for hematopoietic differentiation and should be studied further.

The small differences in sensitivity between K562 cells and CD34+ cells indicate, that differences between these two cell lines are due to changed dynamics and not because of changed topologies.

Bibliography

[1] Fialkow PJ, Jacobson RJ, Papayannopoulou T. Chronic myelocytic leukemia: clonal origin in stem cell common to the granulocyte, erythrocyte, platelet, and monocyte/macrophage. *Am J Med*, 63:125–130, 1977.

[2] Redaelli A, Bell C, Casagrande J, Stephens J, Botteman M, Laskin B, Pashos C. Clinical and epidemiologic burden of chronic myelogenous leukemia. *Expert Rev Anticancer Ther*, 4:85–96, 2004.

[3] Faderl S, Talpaz M, Estrov Z, Kantarjian HM. Chronic myelogenous leukemia: Biology and therapy. *Ann Intern Med*, 131:207–219, 1999.

[4] Stone RM. Optimizing treatment of chronic myeloid leukemia: A rational approach. *The Oncologist*, 9:259–270, 2004.

[5] Sawyers CL. Chronic myeloid leukemia. *N Engl J Med*, 340:1330–1340, 1999.

[6] Rowley JD. A new consistent chromosomal abnormality in chronic myelogeneous leukemia identified by quinacrine fluorescence and giemsa staining. *Nature*, 243:290–293, 1973.

[7] Heisterkamp N, Jenster G, ten Hoeve J, Zovich D, Pattengale PK, Groffen J. Acute leukemia in bcr/abl transgenic mice. *Nature*, 344:251–253, 1990.

[8] Kelliher MA, McLaughlin J, Witte ON, Rosenberg N. Induction of a chronic myelogenous leukemia-like syndrome in mice with v-abl and bcr/abl. *Proc Natl Acad Sci USA*, 87:6649–6653, 1990.

[9] Daley GQ, Van Etten RA, Baltimore D. Induction of chronic myelogenous leukemia in mice by the p210bcr/abl gene of the philadelphia chromosome. *Science*, 247:824–830, 1990.

[10] Silver RT, Woolf SH, Hehlmann R, Appelbaum FR, Anderson J, Bennett C, Goldman JM, Guilhot F, Kantarjian HM, Lichtin AE, Talpaz

M, Tura S. An evidence-based analysis of the effect of busulfan, hydroxyurea, interferon, and allogenic bone marrow transplantation in treating the chronic phase of chronic myeloid leukemia: developed for the american society of hematology. *Blood*, 94:1517–1536, 1999.

[11] Lemischka I. Stem cell dogmas in the genomics era. *Rev Clin Exp Hematol*, 5(1), 2001.

[12] Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA, Lemischka IR. A stem cell molecular signature. *Science*, 298:601–604, 2002.

[13] Clayton S. Hematopoietic stem cells and hematopoiesis. *Cancer Control*, 10:9–16, 2003.

[14] Weissman IL. Stem cells: units of development, units of regeneration, and units in evolution. *Cell*, 100:157–168, 2000.

[15] Osawa M, Hanada K, Hamada H, Nakauchi H. Long-term lymphohematopoietic reconstitution by a single cd34-lownegative hematopoietic stem cell. *Science*, 273:242–245, 1996.

[16] Akashi K, Traver D, Miyamoto T, Weissman IL. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, 404:193–197, 2000.

[17] Kondo M, Akashi K, Weissman IL. Identification of clonogenic lymphoid progenitors in mouse bone marrow. *Cell*, 91:661–672, 1997.

[18] http://www.copewithcytokines.de/cope.cgi?004470.

[19] Graf T. Differentiation plasticity of hematopoietic cells. *Blood*, 99:3089–3101, 2002.

[20] Domen J, Cheshier SH, Weissman IL. The role of apoptosis in the regulation of hematopoietic stem cells: overexpression of bcl-2 increases both their number and repopulation potential. *J Exp Med*, 191:253–264, 2000.

[21] Domen J, Weissman IL. Self-renewal, differentiation or death: regulation and manipulation of hematopoietic stem cell fate. *Mol Med Today*, 5:201–208, 1999.

[22] Orkin SH, Zon LI. Hematopoiesis and stem cells: plasticity versus developmental heterogeneity. *Nat Immunol*, 3:323–328, 2002.

[23] Lemischka IR. The hematopoietic stem cell and its clonal progeny: mechanisms regulating the hierarchy of primitive hematopoietic cells. *Cancer Surv.*, 15:3–18, 1992.

[24] Hackney JA, Charbord P, Brunk BP, Stoeckert CJ, Lemischka IR, Moore KA. A molecular profile of a hematopoietic stem cell niche. *Proc Natl Acad Sci USA*, 99:13061–13066, 2002.

[25] Enver T, Heyworth CM, Dexter TM. Do stem cells play dice ? *Blood*, 92:348–351, 1998.

[26] Enver T, Greaves M. Loops, lineage and leukemia. *Cell*, 94:9–12, 1998.

[27] King AG, Kondo M, Scherer DC, Weissman IL. Lineage infidelity in myeloid cells with tcr gene rearrangement: a latent developmental potential of prot cells revealed by ectopic cytokine receptor signalling. *Proc Natl Acad Sci USA*, 99:4508–4513, 2002.

[28] Kondo M, Scherer DC, Miyamoto T, King AG, Akashi K, Sugamura K, Weissman IL. Cell-fate conversion of lymphoid-committed progenitors by instructive actions of cytokines. *Nature*, 407:383–386, 2000.

[29] Montecino-Rodriguez E, Leathers H, Dorshkind K. Bipotential b-macrophage progenitors are present in adult bone marrow. *Nat Immunol*, 2:83–88, 2001.

[30] Geiger H, Sick S, Bonifer C, Muller AM. Globin gene expression is reprogrammed in chimeras generated by injencting adult hematopoietic stem cells into mouse blastocasts. *Cell*, 93:1055–1065, 1998.

[31] Glimcher LH, Singh H. Transcription factors in lymphocyte development - t and b cells get together. *Cell*, 96:13–23, 1996.

[32] Rothenberg EV, Telfer JC, Anderson MK. Transcription regulation of lympocyte lineage commitment. *Bioessays*, 21:726–742, 1999.

[33] Nagamura-Inoue T, Tamura T, Ozato K. Transcription factors that regulate growth and differentiation of myeloid cells. *Int Rev Immunol*, 20:83–105, 2001.

[34] Yamanaka R, Lekstrom-Himes J, Barlow C, Wynshaw-Boris A, Xanthopoulos KG. Ccaat/enhancer binding proteins are critical components of the transcription regulation of hematopoiesis. *Int J Mol Med*, 1:213–221, 1998.

[35] Tenen DG, Hromas R, Licht JD, Zhang DE. Transcription Factors, Normal Myeloid Development, and Leukemia. *Blood*, 90:489–519, 1997.

[36] Nichols J, Nimer SD. Transcription factors, translocations, and leukemia. *Blood*, 80:2953–2963, 1992.

[37] Shivdasani RA, Orkin SH. The transcriptional control of hematopoiesis [see comments]. *Blood*, 87:4025–4039, 1996.

[38] Benekli M, Baer MR, Baumann H, Wetzler M. Signal transducer and activator of transcription proteins in leukemias. *Blood*, 101:2940–2954, 2003.

[39] Melo JV, Deininger MW. Biology of chronic myelogenous leukemia – signaling pathways of initiation and transformation. *Hematol Oncol Clin North Am*, 18:545–568, 2004.

[40] Ben-Neriah Y, Daley GQ, Mes-Masson AM, Witte ON, Baltimore D. The chronic myelogenous leukemia-specific p210 protein is the product of the bcr-abl hybrid gene. *Science*, 233:212–216, 1986.

[41] Melo JV. The diversity of bcr-abl fusion proteins and their relationship to leukemia phenotype. *Blood*, 88:2375–2384, 1996.

[42] Barnes DJ, Melo JV. Cytogenetic and molecular genetic aspects of chronic myeloid leukaemia. *Acta Haematol*, 108:180–202, 2002.

[43] Calabretta B, Perrotti D. The biology of cml blast crisis. *Blood*, 103:4010–4022, 2004.

[44] http://www.wikipedia.org.

[45] Krebs EG, Beavo JA. Phosphorylation-dephosphorylation of enzymes. *Annu Rev Biochem*, 48:923–959, 1979.

[46] Heinrich R, Neel BG, Rapoport TA. Mathematical models of protein kinase signal transduction. *Mol Cell*, 9:957–970, 2002.

[47] http://www.wordiq.com/.

[48] Hucho F, Buchner K. Signal transduction and protein kinases: the long way from the plasma membrane into the nucleus. *Naturwissenschaften*, 84:281–290, 1997.

[49] Bluthgen N, Herzel H. How robust are switches in intracellular signaling cascades? *J Theor Biol*, 225:293–300, 2003.

[50] Gonze D, Halloy J, Goldbeter A. Robustness of circadian rhythms with respect to molecular noise. *Proc Natl Acad Sci USA*, 99:673–678, 2002.

[51] Ortega F, Acerenza L, Westerhoff HV, Mas F, Cascante M. Product dependence and bifunctionality compromise the ultrasensitivity of signal transduction cascades. *Proc Natl Acad Sci USA*, 99:1170–1175,2002.

[52] Louis M, Becskei A. Binary and graded responses in gene networks. *Sci STKE*, 143:PE33, 2002.

[53] Rebay I. Keeping the receptor tyrosine kinase signaling pathway in check: lessons from drosophila. *Dev Biol*, 251:1–17, 2002.

[54] Swain PS, Elowitz MB, Siggia ED. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci USA*, 99:12795–12800, 2002.

[55] Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A. Regulation of noise in the expression of a single gene. *Nat Genet*, 31:69–73, 2002.

[56] Paulsson J. Summing up the noise in gene networks. *Nature*, 427:415–418, 2004.

[57] Rao CV, Wolf DM, Arkin AP. Control, exploitation and tolerance of intracellular noise. *Nature*, 420:231–237, 2002.

[58] Thattai M, van Oudenaarden A. Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci USA*, 98:8614–8619, 2001.

[59] Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. *Science*, 297:1183–1186, 2002.

[60] Mayani H, Dragowska W, Lansdorp PM. Lineage commitment in human hemopoiesis involves asymmetric cell division of multipotent progenitors and does not appear to be influenced by cytokines. *J Cell Physiol*, 157:579–586, 1993.

[61] Ferrell JE Jr. Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Curr Opin Cell Biol*, 14:140–148, 2002.

[62] Becskei A, Serrano L. Engineering stability in gene networks by autoregulation. *Nature*, 405:590–593, 2000.

[63] Gardner TS, Collins JJ. Neutralizing noise in gene networks. *Nature*, 405:520–521, 2000.

[64] McAdams HH, Arkin A. It's a noisy business! genetic regulation at the nanomolar scale. *Trends Genet*, 15:65–69, 1999.

[65] Cook DL, Gerber AN, Tapscott SJ. Modeling stochastic gene expression: implications for haploinsufficiency. *Proc Natl Acad Sci USA*, 95:15641–15646, 1998.

[66] Hartwell LH, Weinert TA. Checkpoints: controls that ensure the order of cell cycle events. *Science*, 246:629–634, 1989.

[67] Paulsson J, Berg OG, Ehrenberg M. Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. *Proc Natl Acad Sci USA*, 97:7148–7153, 2000.

[68] Thieffry D, Thomas R. Qualitative analysis of gene networks. *Pac Symp Biocomput*, 3:77–88, 1998.

[69] Kohn KW. Molecular interaction map of the mammalian cell cycle control and dna repair system. *Mol Biol Cell*, 10:2703–2734, 1999.

[70] Kauffman SA Homeostasis and differentiation in random genetic control networks. *Nature*, 224:177–178, 1969.

[71] Maki Y, Tominaga D, Okamoto M, Watanabe S, Eguchi Y. Development of a system for the inference of large scale genetic networks. *Pac Symp Biocomput*, 6:446–458, 2001.

[72] Meyers S, Friedland P. Knowledge based simulation of genetic regulation on bacteriophage lambda. *Nucleic Acids Res*, 12:1–9, 1984.

[73] Thomas R, Thieffry D, Kaufman M. Dynamical behaviour of biological regulatory networks: I. biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull Math Biol*, 57:247–276, 1995.

[74] Sanchez L, Thieffry D. A logical analysis of the drosophila gap genes. *J Theor Biol*, 211:115–141, 2001.

[75] Arkin A, Ross J, McAdams HH. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells. *Genetics*, 149:1633-1648, 1998.

[76] Schoeberl B, Eichler-Jonsson C, Gilles ED, Muller G. Computational modeling of the dynamics of the map kinase cascade activated by surface and internalized egf receptors. *Nature*, 20:370–375, 2002.

[77] Ueda HR, Hagiwara M, Kitano H. Robust oscillations within the interlocked feedback model of drosophila circadian rhythm. *J Theor Biol*, 210:401–406, 2001.

[78] De Jong H, Gouze JL, Hernandez C, Page M, Sari T, Geiselmann J. Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bull Math Biol*, 66:301–340, 2004.

[79] Marnellos G, Deblandre GA, Mjolsness E, Kintner C. Delta-notch lateral inhibitory patterning in the emergence of ciliated cells in xenopus: Experimental observations and a gene network model. *Proc Pac Symp Biocomput*, 5:326–337, 2000.

[80] Murphy K, Mian S. Modelling gene expression data using dynamic bayesian networks. *Tech. Rep., University of California*, 1999.

[81] Jensen FV. Bayesian networks basics. *AISB quarterly*, 94:9–22, 1996.

[82] Woolf PJ, Prudhomme W, Daheron L, Daley GQ, Lauffenburger DA. Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. *Bioinformatics*, 21:741–753, 2004.

[83] Cooper GF, Herskovits E. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[84] Huang C,Darwiche A. Inference in belief networks: A procedural guide. *Int J Approx Reasoning*, 11:1–158, 1994.

[85] Friedman N, Linial M, Nachman I, Pe'er D. Using bayesian networks to analyze expression data. *J Comp Biol*, 7:601–620, 2000.

[86] Heckerman D. A tutorial on learning with bayesian networks. *Kluwer Academic Publisher*, 301–354, 1998.

[87] Pe'er D. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17:215–224, 2001.

[88] Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Combining location and expression data for principled discovery of genetic regulatory network models. *Pac Symp Biocomput*, 7:437–449, 2002.

[89] Masys DR. Linking microarray data to the literature. *Nat Genet*, 28:9–10, 2001.

[90] Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S, Miyano S. Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *J Bioinform Comput Biol*, 2:77–98, 2003.

[91] Perrin BE. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19:138–148, 2003.

[92] Zou M, Conzen SD. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21:71–79, 2005.

[93] Cinquin O, Demongeot J. Roles of positive and negative feedback in biological systems. *C R Biol*, 325:1085–1095, 2002.

[94] Cinquin O, Demongeot J. Positive and negative feedback: stricking a balance between necessary antagonists. *J Theor Biol*, 216:229–241, 2002.

[95] Chickering DM. Learning bayesian networks is np-complete. learning from data: Artificial intelligence and statistics. *Springer Verlag*, pages 121–130, 1996.

[96] Zhu J, Emerson SG. Hematopoietic cytokines, transcription factors and lineage commitment. *Oncogene*, 21:3295–3313, 2002.

[97] http://www.biobase.de.

[98] Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mrna expression levels on a genomic scale. *Genome Biol*, 4:117–122, 2003.

[99] Lewin B. Genes vi. *Oxford University Press*, 1997.

[100] Knoth S. Dissertation from Stefan Knoth. *(unpublished)*, 2005.

[101] Robinson RW. Counting unlabeled acyclic digraphs. in c. h. c. little, ed., combinatorial mathematics v, volume 622 of lecture notes in mathematics. *Berlin, Springer*, pages 28–43, 1977.

[102] Pusch L. Personal communication.

[103] Talpaz M, Kantarjian HM, McCredie K, Trujillo JM, Keating MJ, Gutterman JU. Hematologic remission and cytogenetic improvement induced by recombinant human interferon alpha a in chronic myelogenous leukemia. *N Engl J Med*, 314:1065–1069, 1986.

[104] Talpaz *et al.*Talpaz M, Kantarjian H, Kurzrock R, Trujillo JM, Gutterman JU. Interferon-alpha produces sustained cytogenetic responses in chronic myelogenous leukemia: Philadelphia chromosome-positive patients. *Ann Intern Med*, 114:532–538, 1991.

[105] http://www.ebi.ac.uk.

[106] Lawn RM, Gross M, Houck CM, Franke AE, Gray PV, Goeddel DV. DNA sequence of a major human leukocyte interferon gene. *Proc Natl Acad Sci USA*, 78:5435–5439, 1981.

[107] Sattler M, Salgia R, Okuda K, Uemura N, Durstin MA, Pisick E, Xu G, Li JL, Prasad KV, Griffin JD.The proto-oncogene product p120cbl and the adaptor proteins crkl and c-crk link c-abl, p190bcr/abl and p210bcr/abl to the phosphatidylinositol-3' kinase pathway. *Oncogene*, 12:839–846, 1996.

[108] Skorski T, Bellacosa A, Nieborowska-Skorska M, Majewski M, Martinez R, Choi JK, Trotta R, Wlodarski P, Perrotti D, Chan TO, Wasik MA, Tsichlis PN, Calabretta B. Transformation of hematopoietic cells by bcr/abl requires activation of a pi3k/akt-dependent pathway. *EMBO J*, 16:6151–6161, 1997.

[109] Gesbert F, Sellers WR, Signoretti S, Loda M, Griffin JD. Bcr/abl regulates expression of the cyclin-dependent kinase inhibitor p27kip1 through the phosphatidylinositol 3-kinase/akt pathway. *J Biol Chem*, 275:39223–39230, 2000.

[110] Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[111] Chang L, Karin M. Mammalian map kinase signalling cascades. *Nature*, 410:37–40, 2001.

[112] Franke TF, Kaplan DR, Cantley LC. Pi3k: Downstream aktion blocks apoptosis. *Cell*, 88:435–437, 1997.

[113] Varticovski L, Daley GQ, Jackson P, Baltimore D, Cantley LC Activation of phosphatidylinositol 3-kinase in cells expressing abl oncogene variants. *Mol Cell Biol*, 11:1107–1113, 1991.

[114] Skorski T, Kanakaraj P, Nieborowska-Skorska M, Ratajczak MZ, Wen SC, Zon G, Gewirtz AM, Perussia B, Calabretta B. Phosphatidylinositol-3 kinase activity is regulated by bcr/abl and is required for the growth of philadelphia chromosome-positive cells. *Blood*, 86:726–736, 1995.

[115] Myklebust JH, Blomhoff HK, Rusten LS, Stokke T, Smeland EB. Activation of phosphatidylinositol 3-kinase is important for erythropoietin-induced erythropoiesis from cd34(+) hematopoietic progenitor cells. *Exp Hematol,*, 30:990–1000, 2002.

[116] http://caroll.vjf.cnrs.fr/cancergene/home.html.

[117] http://www.ncbi.nlm.nih.gov.

[118] http://au.expasy.org.

Abbreviations

Abbreviations of genes and molecules are used as in the Tranpath database.

- : The activity of the molecule gets decreased or is already zero.

+ : The activity of the molecule gets increased or is already maximal.

0 : The activity of the molecule is not changed

Abl : Abelson tyrosine kinase

ABSENT : Affymetrix Suite indicates that gene is not expressed.

Acp5 : Acid phosphatase 5

Akt : Protein Kinase B, PKB

AML1 : Acute Myeloid Leukemia 1; alternatively: CBFA2 (Core Binding Factor A2), PEBPaB (Polyomavirus Enhancer Binding Protein aB)

AQP 7 : Aquaporin 7

AraC : Cytosin Arabinosid

ATF-2 : Activating Transcription Factor 2

B : B lymphocyte

Bcl-2 : B-cell lymphoma-2

BCR : Breakpoint Cluster Region (gene or protein)

β**ARK-1 :** Beta-Adrenergic Receptor Kinase 1; alternatively:

Bmx : Bone marrow kinase

CAD : Caspase-Activated DNAse

CaMKIV : Ca2+/calmodulin-dependent Kinase IV

cAMP : cyclic Adenosin Monophospate

Cat-2 : Cool-associated tyrosine-phosphorylated protein 2

CBP : CREB-Binding Protein

CCR5 : CC chemokine Receptor 5

CD80 : T Lymphocyte Activation Antigen CD80 Precursor

Cdc42 : Cell division cycle protein 42

Cdk : Cyclin-dependent kinase

cGMP : cyclic Guanosine-Monophosphate

Ci : Cubitus interruptus

CLP : Common Lymphoid Progenitor

CML : Chronic myelogenous leukaemia

CMP : Common Myeloid Progenitor

CMP : Common Multipotent Progenitor

Cos2 : Costal2

Co-Smads : Common mediator Smads

COX2 : Cyclooxygenase 2;

Csk : C-terminal Src-Kinase

CSN2 : Casein beta

CTLA-4 : Cytotoxic T Lymhocyte Antigen 4

DAG : Diacylglycerol

DBN : Dynamic Bayesian Network

DecrExpr : Affymetrix Suite indicates that expression of the gene is decreased.

E : Erythrocyte

E. coli : Escherichia Coli

Eb : Erythroblast

EGFR : Epidermal Growth Factor Receptor

Egr3 : Early Growth Response 3

eNOS : Endothelial Nitric Oxide Synthase

Eos : Eosinophil

EPO : Erythropoietin

EpoR : Erythropoietin Receptor

ER-α : Estrogen Receptor-alpha

ERbB2 : v-Erb-b2 erythroblastic leukaemia viral oncogene homolog 2

ERK : Extracellular signal Regulated Kinase

FAK1 : Focal Adhesion Kinase 1

FcgammaRI : FcgammaReceptor I

FCγRIIB : FcgammaReceptor IIB

fMLP-R : formyl-Methionyl-Leucyl-Phenylalanine Receptor

FTase : Farnesyltransferase

Fu : Fused

G : neutrophil Granulocyte

G6PC : Glucose-6-Phosphatase

GM-CSF : Granulocyte-Macrophage Colony Stimulating Factor

GMP : Granulocyte/Macrophage Progenitor

GPCR : G-Protein Coupled Receptor

Grb-2 : Growth-factor Receptor-Binding Protein 2

GSK3 : Glycogen Synthase Kinase 3

HePTP : Hematopoietic Protein Tyrosine Phosphatase

HLA-D : Human Leukocyte Antigen D

HNF-6 : Hepatocyte Nuclear Factor 6

HNF4A : Hepatocyte Nuclear Factor 4 alpha

HSC : Hematopoietic Stem Cell

HSP27 : Heat Shock Protein 27

IAP : Integrin Associated Protein

ICAD : Inhibitor of CAD

ICAM-1 : Intercellular Adhesion Molecule 1

IFN : Interferon

IFN-γ : Interferon gamma

IFNRI : Interferon Receptor I

IFNRII : Interferon Receptor II

IKK : IkB Kinase

IL-2 : Interleukin 2

IL-2R : Interleukin 2 Receptor

IL-6R : Interleukin 6 Receptor

ILK : Integrin Linked Kinase

IncrExpr : Affymetrix Suite indicates that expression of the gene is increased.

ins-1 : Insulin 1

InsR : Insulin Receptor

IP-10 : Interferon gamma inducible Protein

IP_3 : Inositol-1,4,5-Trisphosphate

IP3R : Inositol-1,4,5-Trisphosphate Receptor

ISG15 : Interferon Stimulated Gene 15

Jak : Janus Kinase

Jak 2: Janus Kinase 2

JNK3 : c-Jun N-Terminal Kinase 3

LXR : Liver X Receptor

Mac : Macrophage

MAP Kinase : Mitogen Activated Pathway Kinase

MCP-1 : Monocyte Chemoattractant Protein 1

MEFV : Mediterranean fever

Meg : Megakaryocyte

MEK : Mitogen Activated Extracellular Protein Kinase

MEP : Megakaryocyte Erythrocyte Progenitor

Mst : Mammalian Ste20-like Kinase

Nck-2 : Non-Catalytic Region of Tyrosine Kinase Adaptor Protein 2

NF-AT : Nuclear Factor of Activated T-Cells

NHERF-1 : Na+/H+ Exchange Regulatory Factor

NK : Natural Killer cell

NO : Nitric Oxide

NP-complete : Non-deterministic Polynomial time complete

NQO1 : NAD(PH) dehydrogenase quinone 1

NQO1 : Ordinary Differential Equation

Oxt : Oxytocin

P : Platelet

p70S6K : Ribosomal Protein S6 Kinase

PARP : Poly(ADP-ribose)Polymerase

PDK1 : 3'-Phospphatinositide Dependent Kinase 1

PEP : PEST Domain-Enriched Tyrosine Phosphatase

Ph : Philadelphia chromosome

Pi3K : Phosphatidylinositol 3-Kinase

PIAS : Protein Inhibitor of Activated STAT

PIP 2: Phosphatidyl Inositol-4,5-bisphosphate

PIP3 : Phosphatidylinositol-(3,4,5)triphosphat

PITSLRE : p34:cdc2-Related protein kinase

PKA : Protein Kinase A

PKC : Protein Kinase C

PKR : Protein Kinase RNA Regulated

PLC : Phospholipase C

PLCγ : Phospholipase C gamma

PR : Progesterone Receptor

pRb : Retinoblastome Associated Protein

PRESENT : Affymetrix Suite indicates that gene is expressed.

PRK-2 : PKC Related protein Kinase 2

Ps2 : Preselinin 2

PTMA : Prothymosine Alpha

R-Smads : Receptor regulated Smads

RB : Retinoblastoma 1

Rbp : Retinil Binding Protein 4

Ref-1 : Redox Factor 1

RICS : RhoGAP involved in the beta catenin N cadherin and NMDA receptor signaling

RSK : Ribosomal S6 Kinase

Sfpi1 : SFFV Proviral Integration 1

SHIP : SH-2-Containing Inositol-5-Phosphatase

SHP-1 : SH-2-Containing Phosphatase 1

SLC9A3R1 : Solute Carrier Family 9 isoform 3 Regulatory factor 1

SLP-65 : SH-2-Domain Containing Leukocyte Adaptor of 65 kDa

SLP-76 : SH-2-Domain Containing Leukocyte Adaptor of 76 kDa

SOCS-3 : Suppressor of Cytokine Signalling 3

SOS : Son of sevenless

SPRR1B : Small Proline-Rich Protein 1B

STAT3 : Signal Transducer and Activator of Transcription 3

STAT5 : Signal Transducer and Activator of Transcription 5

T : T lymphocyte

Tf : Transferrin

TGFα : Tranforming Growth Factor Alpha

TIMP-1 : Tissue Inhibot of Metalloproteinase 1

TNF : Tumor Necrosis Factor

TNF-α : Tumor Necrosis Factor Alpha

TNFR : Tumor Necrosis Factor Receptor

TPO : Thrombopoietin

uk : The activity of the molecule cannot be calculated exactly, therefore it is considered to be 50% + and 50 % -.

Acknowledgments

I would like to thank all people who supported me during the time of this work. I am especially indepted to:

Prof. Dr. Stefan Wölfl, for providing the opportunity to work in his group on a very interesting topic.

PD Dr. Reinhard Guthke and Martin Hoffmann from the Department of Applied Microbiology at the Hans Knöll Institute for Natural Products Research for important suggestions and discussions.

Stefan Knoth for doing the required experiments.

Torsten Kroll, who shared the office with me, for helping me in solving many organisational problems and interesting and funny discussions.

Florian Centler and Dr. Peter Dittrich and other members of the Bio systems Analysis group at the University Jena, for serious suggestions and funny moments.

The members of the Junior Research Group – Theoretical Systems Biology and here especially Dr. Andreas Beyer for discussions.

All the members of my group for the nice working atmosphere.

Special thanks to Altaf Ahmad Dar for friendship and for having an open ear for all my problems not only related with work.

My girlfriend Cornelia Heller for reading my thesis critically and her constant support.